

Evaluation Methodologies for Visual Information Retrieval and Annotation

Dissertation

zur Erlangung des akademischen Grades
Doktoringenieur (Dr.-Ing.)

vorgelegt der Fakultät für Elektrotechnik und Informationstechnik
der Technischen Universität Ilmenau

von Dipl.-Inf. Stefanie Nowak
geboren am 16. November 1980 in Paderborn

Tag der Einreichung: 05.07.2011
Tag der wissenschaftlichen Aussprache: 16.12.2011

Gutachter:

1. Univ-Prof. Dr.-Ing. Dr. rer. nat. h.c. mult. Karlheinz Brandenburg
2. Prof. Dr. rer. nat. Stefan Rüger
3. Prof. Dr. rer. nat. Henning Müller

URN: urn:nbn:de:gbv:ilm1-2011000499

Danksagung

Diese Arbeit ist nur durch die Hilfe und großartige Unterstützung vieler Personen möglich gewesen. Ich möchte mich ganz besonders bei meinem Doktorvater Prof. Brandenburg bedanken. Sie haben mich meinen Weg in der Forschungsgemeinschaft gehen lassen, mir ermöglicht an den relevanten Konferenzen teilzunehmen und auch keine Unterstützung gescheut, wenn es um Vorhaben wie Forschungsaufenthalte oder flexible Arbeitsstandorte ging. Darüber hinaus habe ich für sämtliche Fragen ein offenes Ohr gefunden. Prof. Rüger, vielen Dank für Ihre tolle Unterstützung und die Möglichkeit in Ihrer Gruppe arbeiten zu dürfen. Die Zeit am KMI ist mir als ein Highlight meiner Doktorarbeitsphase in Erinnerung geblieben. Ich durfte viel von Ihnen lernen, sowohl Fachliches als auch deutsch-englische Kultur. Prof. Müller, vielen Dank für das große Vertrauen und die geduldige Beantwortung sämtlicher Fragen über Evaluationsbenchmarks zu jeder Tages- und Nachtzeit. An der Organisation des ImageCLEFs mitwirken zu dürfen und selbständig Entscheidungen treffen zu dürfen, hat mich wachsen lassen und mir all die Jahre äußerst viel Freude bereitet. Ein großer Dank geht an das Projekt THESEUS und den DAAD für die finanzielle Unterstützung bei Konferenzen und Reisen, sowie die Möglichkeit aktiv den ImageCLEF mitgestalten zu können.

Dem IDMT, der Doktorandengruppe vom Institut für Medientechnik und insbesondere der Metadaten Abteilung des IDMT möchte ich aufrichtig für die Unterstützung und fruchtbare Zusammenarbeit über die ganzen Jahre danken. Die institutsweite Photo-Annotationsaktion war unglaublich. Vielen Dank für die Zeit und Energie, und dass ihr euch alle mit diesem Thema (oftmals in der Freizeit) auseinander gesetzt habt. Ein großer Dank geht an Holger Großmann, Christian Dittmar, Uwe Kühhirt und Peter Dunker. Ihr habt mir ermöglicht, am Projekt Dissertation dran zu bleiben und dieses auch im manchmal turbulenten Projektalltag weiter verfolgen zu können. Dabei vielen Dank, Uwe, für die Zeit, die du mir insbesondere am Ende meiner Arbeit freigeschaufelt hast. Peter, danke für dein Engagement, deine Ideen und die große Unterstützung zu den Anfangszeiten von IDMT@ImageCLEF. Danke an meine Diplomanden, insbesondere an Karolin Nagel und Tilman Sieweke: Eure Arbeit hat mit zum Gelingen dieser Arbeit beigetragen. Nicht zuletzt einen riesigen Dank an die zahlreichen Korrekturleser Jakob Abeßer, Thomas Henning, Judith Liebetrau, Ania Lukashevich, Enrico May, Karolin Nagel, Gerhard Nowak, Ronny Padoschek, Julia Reinbach und Sebastian Schneider für eure Hinweise und Anmerkungen.

Ania, vielen Dank für deine unendliche Geduld, Dinge zu erklären, sorgfältigst Korrektur zu lesen oder sich über Evaluation und Machine Learning im Allgemeinen auszutauschen. Ich werde die nächtlichen Skype-Sessions zu Paper Deadlines in guter Erinnerung behalten und freue mich insbesondere, dass die Arbeit sich am Ende ausgezahlt hat. Ich vermisse unser Büro. Judith, vielen lieben Dank für deine Unterstützung in jeglicher Hinsicht, sei es die Bewahrung eines kühlen Kopfes in heißen Zeiten, die Umorganisation von Dingen, so dass es am Ende alles passte, Ideen für ImageCLEF, Korrekturen, Annotationen oder die lustige Zeit bei nicht enden wollenden Feierabendbieren. Deine Reiseführer-Kompetenzen habe ich immer gerne in Anspruch genommen und hoffe, dass wir noch ein paar Gelegenheiten bekommen, zusammen unterwegs zu sein. Dominik, ich habe unsere gemeinsamen Heimfahrten genossen und danke wegen der vielen Tipps, die du auf Lager hattest, sei es über Verfahren zur nutzerbasierten Evaluation, Gliederungsprobleme oder Fragen organisatorischer Natur.

An die Metadatentruppe: Die Zeit mit euch war superschön. Ich habe es genossen, mit euch nicht nur zusammen arbeiten zu dürfen, sondern auch diverse Freizeitaktivitäten zu gestalten. Meine persönlichen Highlights sind Fahrradfahrten im Hochwasser, sämtliche Havanna Abende, Bastelaktionen, Fußball gucken bei Daniel, das Highfield, und nicht zu vergessen die Sommercons.

Guys from KMI, thanks so much for your love of life, your happiness, and your support. After one week in MK, I felt like I have been there for a long time. Miriam, thanks personally and professionally. I will never forget all efforts you did and the smile on your face no matter how stressful life was. Thomas, you made life so nice and straightforward with your kindness and your way of taking care of everybody and everything. Thank you. Ainhoa, I enjoyed our collaboration and our coffee breaks very much. I hope it was not the last time working together. Thank you Ivana, Marco, Serge, Alba, Vanessa, Philippe, Suzanne, and Georgios for the great time we spent.

1000 Dank an dich, Christoph, für deine Geduld, deine fachliche und ideelle Unterstützung, sowie die vielen aufmunternden Worte auf dem langen Weg auf dem diese Doktorarbeit entstanden ist. Deine Hilfe, deine Frustrationstoleranz und die stundenlangen Diskussionen darüber welcher Weg der Bessere ist, haben mir geholfen Täler zu überwinden und Lücken aufzudecken. Darüberhinaus hast du sämtliche Ideen meinerseits über Dinge wie Auslandsaufenthalte oder Konferenzen nicht nur akzeptiert, sondern sie großartig unterstützt und oftmals meine eigenen Zweifel daran besiegt. Dankeschön.

An meine Familie: Danke für eure Liebe und bedingungslose Unterstützung.

Abstract

Performance assessment plays a major role in the research on Information Retrieval (IR) systems. Starting with the Cranfield experiments in the early 60ies, methodologies for the system-based performance assessment emerged and established themselves, resulting in an active research field with a number of successful benchmarking activities. With the rise of the digital age, procedures of text retrieval evaluation were often transferred to multimedia retrieval evaluation without questioning their direct applicability. This thesis investigates the problem of system-based performance assessment of annotation approaches in generic image collections. It addresses three important parts of annotation evaluation, namely user requirements for the retrieval of annotated visual media, performance measures for multi-label evaluation, and visual test collections. Using the example of multi-label image annotation evaluation, I discuss which concepts to employ for indexing, how to obtain a reliable ground truth to moderate costs, and which evaluation measures are appropriate. This is accompanied by a thorough analysis of related work on system-based performance assessment in Visual Information Retrieval (VIR). Traditional performance measures are classified into four dimensions and investigated according to their appropriateness for visual annotation evaluation. One of the main ideas in this thesis adheres to the common assumption on the binary nature of the score prediction dimension in annotation evaluation. However, the predicted concepts and the set of true indexed concepts interrelate with each other. This work will show how to utilise these semantic relationships for a fine-grained evaluation scenario. Outcomes of this thesis result in a user model for concept-based image retrieval, a fully assessed image annotation test collection, and a number of novel performance measures for image annotation evaluation.

Zusammenfassung

Die automatisierte Evaluation von Informations-Retrieval-Systemen erlaubt Performanz und Qualität der Informationsgewinnung zu bewerten. Bereits in den 60er Jahren wurden erste Methodologien für die system-basierte Evaluation aufgestellt und in den Cranfield Experimenten überprüft. Heutzutage gehören Evaluation, Test und Qualitätsbewertung zu einem aktiven Forschungsfeld mit erfolgreichen Evaluationskampagnen und etablierten Methoden. Evaluationsmethoden fanden zunächst in der Bewertung von Textanalyse-Systemen Anwendung. Mit dem rasanten Voranschreiten der Digitalisierung wurden diese Methoden sukzessive auf die Evaluation von Multimediaanalyse-Systeme übertragen. Dies geschah häufig, ohne die Evaluationsmethoden in Frage zu stellen oder sie an die veränderten Gegebenheiten der Multimediaanalyse anzupassen.

Diese Arbeit beschäftigt sich mit der system-basierten Evaluation von Indizierungssystemen für Bildkollektionen. Sie adressiert drei Problemstellungen der Evaluation von Annotationen: Nutzeranforderungen für das Suchen und Verschlagworten von Bildern, Evaluationsmaße für die Qualitätsbewertung von Indizierungssystemen und Anforderungen an die Erstellung visueller Testkollektionen. Am Beispiel der Evaluation automatisierter Photo-Annotationsverfahren werden relevante Konzepte mit Bezug zu Nutzeranforderungen diskutiert, Möglichkeiten zur Erstellung einer zuverlässigen Ground Truth bei geringem Kosten- und Zeitaufwand vorgestellt und Evaluationsmaße zur Qualitätsbewertung eingeführt, analysiert und experimentell verglichen. Traditionelle Maße zur Ermittlung der Performanz werden in vier Dimensionen klassifiziert. Evaluationsmaße vergeben üblicherweise binäre Kosten für falsche Annotationen. Diese Annahme steht im Widerspruch zu der Natur von Bildkonzepten. Das gemeinsame Auftreten von Bildkonzepten bestimmt ihren semantischen Zusammenhang und von daher sollten diese auch im Zusammenhang auf ihre Richtigkeit hin überprüft werden. In dieser Arbeit wird aufgezeigt, wie semantische Ähnlichkeiten visueller Konzepte automatisiert abgeschätzt und in den Evaluationsprozess eingebracht werden können. Die Ergebnisse der Arbeit inkludieren ein Nutzermodell für die konzeptbasierte Suche von Bildern, eine vollständig bewertete Testkollektion und neue Evaluationsmaße für die anforderungsgerechte Qualitätsbeurteilung von Bildanalyse-Systemen.

Contents

1. Introduction	1
1.1. Motivation	1
1.2. Research objectives	2
1.3. Contributions	3
1.4. Outline of the thesis	5
1.5. Publications	6
I. Context and related work	9
2. Visual Information Retrieval and annotation	13
2.1. Introduction to image annotation	13
2.2. The basic image annotation and evaluation pipeline	15
2.2.1. Descriptors	16
2.2.2. Classifier	17
2.2.3. Fusion techniques and post-processing	18
2.3. Explicit semantics in Visual Information Retrieval	19
2.4. Related work on multi-label image annotation	20
2.5. Summary	22
3. Evaluation of visual Information Retrieval and annotation	23
3.1. History of experimental evaluation	23
3.2. Test collections for VIR	25
3.2.1. Requirements	25
3.2.2. Visual concepts	26
3.2.3. Relevance assessment	27
3.2.4. Databases	29
3.3. Evaluation measures	29
3.3.1. General notations	30
3.3.2. Concept-based evaluation measures for label set predictions	31
3.3.3. Concept-based evaluation measures for ranked assignments	32
3.3.4. Example-based evaluation measures for label set predictions	33
3.3.5. Example-based evaluation measures for ranked assignments	34
3.3.6. Significance tests	35
3.4. Benchmarks for multimedia retrieval	35
3.5. Summary	37

4. Related work on evaluation methodologies for image annotation	39
4.1. Visual concepts for image annotation	39
4.1.1. Discussion	40
4.2. Relevance assessment adopting crowdsourcing	41
4.2.1. Crowdsourcing at Amazon Mechanical Turk	41
4.2.2. Quality assurance of relevance judgements	42
4.2.3. Crowdsourcing for image annotation	42
4.2.4. Discussion	43
4.3. Performance measures for image annotation evaluation	43
4.3.1. Performance measures in image annotation literature	44
4.3.2. Performance measures in VIR benchmarks	46
4.3.3. Discussion	48
4.4. Performance measures considering semantics	48
4.4.1. Hierarchical measures for uni-label classification evaluation	48
4.4.2. Hierarchical measures for multi-label classification evaluation	49
4.4.3. Ontology-based performance measures	50
4.4.4. Comparison	50
4.4.5. Discussion	50
4.5. Semantic relatedness measures	52
4.5.1. Thesaurus-based approaches	52
4.5.2. Wikipedia-based approaches	53
4.5.3. Web-based approaches	53
4.5.4. Discussion	55
4.6. Evaluating evaluation measures	55
4.6.1. Rank correlation	56
4.6.2. Stability analysis	57
4.6.3. Discussion	58
4.7. Summary	58
II. Test collections for multi-label image annotation evaluation	61
5. Topic definition: Which concepts do users prefer to organise photo sets?	65
5.1. Motivation	65
5.2. User studies on photo organisation	66
5.3. How do people describe and categorise images?	67
5.4. User tag analysis	68
5.5. Definition of visual concepts and the Photo Tagging ontology	69
5.5.1. Photo Tagging ontology: V 1.0	70
5.5.2. Photo Tagging ontology: V 2.0	73
5.5.3. Photo Tagging ontology: V 3.0	73
5.6. Comparison to visual concepts of other test collections	74
5.7. Summary	78
6. Relevance assessment: Concept subjectivity and inter-annotator agreement	79
6.1. Motivation	79
6.2. Experimental setup	80
6.2.1. Collecting data of expert annotators	80
6.2.2. Collecting data of non-expert annotators	81

6.3.	Evaluation design	84
6.3.1.	Accuracy for agreement assessment	86
6.3.2.	Rank correlation as a measure of inter-rater agreement	86
6.3.3.	Kappa statistics	86
6.4.	Results	86
6.4.1.	Experiment 1: Agreement analysis among experts	87
6.4.2.	Experiment 2: System ranking with expert annotations	88
6.4.3.	Experiment 3: Agreement analysis between experts and non-experts	89
6.4.4.	Experiment 4: Ranking with non-expert annotations	91
6.5.	Discussion	92
6.6.	Relevance from weakly labelled corpora	93
6.7.	Summary	93
III.	Novel evaluation methodologies for multi-label annotation evaluation	95
7.	A fine-grained evaluation measure for multi-label annotation evaluation	99
7.1.	Motivation	99
7.2.	User requirements for a new multi-label evaluation measure	100
7.2.1.	User model for annotation evaluation	100
7.2.2.	Requirements on multi-label evaluation measures	101
7.3.	Ontology Score: A fine-grained multi-label classification performance measure	101
7.4.	Exemplary comparison of OS to hierarchical measures	105
7.5.	The effectiveness of performance measures in image annotation	107
7.5.1.	Choice of configurations for the case study	107
7.5.2.	Evaluation measures	108
7.5.3.	Study 1: Influence of overannotation on performance score	108
7.5.4.	Study 2: General characteristics of example- and concept-based measures	111
7.5.5.	Discussion	115
7.6.	Summary	116
8.	The effect of semantic relatedness measures on annotation evaluation	117
8.1.	Motivation	117
8.2.	Semantic relatedness measures reconsidered	118
8.3.	Study setup on evaluation using semantic relatedness	119
8.3.1.	Evaluation framework	119
8.3.2.	Data	120
8.3.3.	Configurations	120
8.4.	Study results	121
8.4.1.	Ranking results	121
8.4.2.	Results of stability experiment	123
8.4.3.	Discussion	125
8.5.	Flickr as source for semantic relatedness	125
8.5.1.	Correlation of human judgements to semantic relatedness measures	126
8.5.2.	Semantic relatedness of visual concepts	126
8.6.	Summary	129
9.	Exploiting semantics in a multi-label performance measure for ranked predictions	131
9.1.	Motivation	131
9.2.	The need for yet another evaluation measure	132

9.3. Semantic R-Precision	134
9.4. Experimental work	135
9.4.1. Results of the correlation experiment	136
9.4.2. Results of the stability experiment	140
9.4.3. Comparison of performance scores between SR-Precision and R-Precision	141
9.4.4. Discussion	142
9.5. How close is the user model to real user needs?	143
9.6. Summary	145
10. Image annotation evaluation in ImageCLEF	147
10.1. Motivation	147
10.2. Visual concept detection and annotation task	148
10.2.1. Evaluation objectives in 2009	148
10.2.2. Evaluation objectives at the ICPR contest 2010	149
10.2.3. Evaluation objectives in 2010	149
10.3. Test collection	149
10.3.1. Relevance assessment in 2009	150
10.3.2. Relevance assessment in 2010	151
10.4. Evaluation measures	152
10.5. Submission	152
10.6. Participation	153
10.7. Results	155
10.7.1. Results of the VCDT in ImageCLEF 2009	155
10.7.2. Results of the VCDT at ICPR 2010	157
10.7.3. Results of the VCDT in ImageCLEF 2010	158
10.7.4. Comparison of the three configurations	160
10.8. Evolvement of concept detection performance	162
10.8.1. Evolvement from ImageCLEF 2009 to ICPR 2010	162
10.8.2. Evolvement from ICPR 2010 to ImageCLEF 2010	165
10.8.3. Discussion	165
10.9. Outlook	166
10.10. Summary	166
IV. Conclusions	167
11. Conclusion and future work	169
11.1. Summary	169
11.2. Contributions	170
11.3. Limitations	173
11.4. Directions on future work	176
V. Appendix	I
A. List of visual concepts in ImageCLEF	III
A.1. Visual concepts in ImageCLEF 2009 and ICPR 2010	III
A.2. Visual concepts in ImageCLEF 2010	IV
A.3. Visual concepts in ImageCLEF 2011	V

Contents

B. Theses	VII
Acronyms	IX
Bibliography	XIII
Own publications	XXXV

List of Tables

4.1.	Relation of performance measures to evaluation dimensions	45
4.2.	Consideration of evaluation dimensions in annotation approaches	45
4.3.	Categorisation of performance measures	45
4.4.	Overview of performance measures utilised in annotation benchmarks	47
4.5.	Comparison of hierarchical classification approaches	51
5.1.	Grouping of visual concepts in categories	70
5.2.	Comparison of PTO concepts with other collections	75
6.1.	Comments from turkers	84
6.2.	Annotator accuracy	87
6.3.	Correlation for OS with varying ground truth	89
7.1.	Results of hierarchical performance measures	105
7.2.	Largest variances in measure scores for VCDT results	108
7.3.	Overview of performance measures	109
8.1.	Correlation of different semantic relatedness measures	122
8.2.	Correlation results for stability experiment	124
8.3.	Human judgement on semantic relatedness	127
8.4.	Correlations to Gracia and Mena dataset	127
8.5.	Correlations to visual concepts dataset	129
9.1.	Average correlation of image rankings	137
9.2.	Correlation between overall rankings	137
9.3.	Correlation of runs evaluated with altered and original ground-truth	140
9.4.	Comparison of concept sets	144
9.5.	Results on user model validation	145
9.6.	Expectations of turkers on image retrieval results	146
10.1.	Number of photos and concepts used in the ImageCLEF cycles	150
10.2.	Participation in the VCDT	153
10.3.	Summary of the results for the VCDT in 2009.	155
10.4.	Overview of concepts and results for the VCDT in 2009 and at ICPR 2010	156
10.5.	Results of the ICPR Photo Annotation task	157
10.6.	Results for the VCDT in 2010 per group	158
10.7.	Results per concept of the VCDT 2010 for the visual configuration	159
10.8.	Results per concept of the VCDT 2010 for the textual configuration	159
10.9.	Results per concept of the VCDT 2010 for the multi-modal configuration	159
10.10.	Best annotation performance per concept in the VCDT 2010	161
10.11.	Best annotation performance per concept at ICPR and in ImageCLEF 2010	164

List of Tables

A.1. Visual concepts in ImageCLEF 2009 and at ICPR 2010.	III
A.2. Visual concepts in ImageCLEF 2010.	IV
A.3. Visual concepts in ImageCLEF 2011.	V

List of Figures

2.1. Visual impression of concepts	14
2.2. The basic annotation process	15
3.1. Evaluation cycle for evaluation benchmarks	36
4.1. Dimensions of evaluation measures	44
4.2. Frequency of performance measures	46
6.1. Image annotation tool	81
6.2. HIT template	82
6.3. Statistics of rejected images at MTurk	83
6.4. Agreement among annotators per concept	88
6.5. Kappa statistics for disjunctive concepts	90
6.6. Kappa statistics for optional concepts	91
7.1. Illustration of user model	101
7.2. Visualisation of a fragment of the ontology for image annotation	102
7.3. Schematic representation of the components of the OS.	103
7.4. Fragment of the Photo Tagging Ontology	104
7.5. Schematic representation of the ontology fragment	106
7.6. Influence of LD on performance measure scores	110
7.7. Results of concept-based and example-based evaluation measures	112
7.8. Results for all run configurations	114
8.1. Schematic representation of the evaluation framework	120
8.2. Dendrogram for semantic relatedness measures	123
8.3. MTurk survey about visual concept correlations	128
9.1. Ranking of runs evaluated with R-Precision and F-measure	133
9.2. Dendrogram of example-based performance measures	138
9.3. Scatter plots for evaluation scores	139
9.4. Comparison of R-Precision and SR-Precision scores	141
9.5. Scatter plots for image scores	142
10.1. Example images of VCDT in 2009	150
10.2. MTurk HIT template for vehicle concepts	152
10.3. Frequency of labels compared to iAP scores	160
10.4. Concept occurrences for ImageCLEF 2009 and ICPR 2010	163
10.5. Images with the lowest detection rate	165
A.1. Map of PTO of ImageCLEF 2011	VI

1. Introduction



This chapter comprises a general overview of the thesis. Especially, it provides the motivation for the research conducted in this work and details the contributions. Section 1.1 outlines the research problem. Section 1.2 poses research questions on the topic of evaluation methodologies for image annotation. Next, Section 1.3 summarises the contributions and outcomes of my work on performance assessment for image annotation. Finally, Section 1.4 outlines the structure of this document, while Section 1.5 lists the publications that relate to the work in this thesis.

1.1. Motivation

With the increasing amount of digital information on the Web and on personal computers, the need for systems that are capable of automated indexing, searching, and organising multimedia documents incessantly grows. Automated systems have to retrieve information with a high performance in order to be accepted by industry and end-users. The performance assessment is mostly conducted in a system-based manner by querying the retrieval algorithm with predefined topics and evaluating the retrieved answers based on human judgements. Often, multimedia retrieval systems are evaluated on different test collections, which makes the comparison of retrieval performance impossible. To cope with this issue, multiple benchmarking initiatives focus on the performance assessment of multimedia systems with standardised test collections and tasks, such as in ImageCLEF (Müller et al. (2010)), PASCAL Visual Object Classes (VOC) (Everingham et al. (2010)), and TREC Video retrieval evaluation (TRECVID) (Smeaton et al. (2006)).

The design and setup of evaluation methodologies need to take several aspects into account. The acquisition of a realistic test collection has to be carried out with respect to the evaluation goal, including the definition of topics and the assessment of relevance. Usually, the relevance assessment is the most time- and cost-consuming part of the evaluation. Methods such as pooling try to counteract this issue by reducing the amount of work for the expert judges. Further, the definition of user preferences for query topics and concepts plays a key role. It must be assured that systems are capable of indexing concepts which the users are interested in to add value to an

application. Finally, it must be defined which characteristics determine the “retrieval quality” of a retrieval result. The retrieval quality of general information from the Web can be characterised by the precision and diversity of the retrieved media items with respect to the search query, while it is less important to provide a complete list of all matching documents. In other scenarios in which closed databases are indexed (e.g., patent search in logos, text, or sketches), it might be of great importance to retrieve all potential relevant documents.

Consequently, one key issue in research on multimedia quality assessment is the investigation of appropriate evaluation methodologies. It is necessary for measures to be unbiased and stable against random annotations, and they must reflect the user requirements on multimedia retrieval systems, even when they are evaluated in a system-based manner. Traditional IR performance measures focus on different aspects of quality. Precision and Recall reflect exactly the scenarios outlined above, while the F-Measure combines both characteristics in one score. Other evaluation measures, such as Mean Average Precision (MAP), Area-Under-Curve (AUC), or Equal Error Rate (EER), are capable of incorporating confidences of the retrieval systems in their score.

These traditional performance measures have emerged from research on text retrieval evaluation and are also commonly applied to visual retrieval and annotation systems. However, regarding the evaluation of visual annotation systems, other means of evaluation might be more appropriate. Today’s annotation systems are able to assign a number of semantic concepts to each media item. This approach is a generalisation of the categorisation approach that assigns one single concept to a media object out of a number of categories. As a result, an annotation system predicts a *set of concepts* for each media item. Usually, the MAP or inferred variants are adopted in annotation evaluation and the system prediction is evaluated for each concept in isolation. However, these traditional performance measures do not consider the semantic relatedness between the set of predicted concepts and the ground truth, which might provide important information in an annotation task. Imagine two indexing systems that index an image with the concept set {flower, rose, garden} and the concept set {flower, garden, architecture}. These predictions are judged equally by traditional evaluation measures with respect to the ground truth set {flower, garden}. However, the concept rose is semantically related to the concepts garden and flower and is indeed a specialisation of flower, while the concept architecture does not have such a high semantic relatedness to the other concepts. This issue should be considered in annotation evaluation, as it provides an adequate means to distinguish among the quality of annotation systems. Instead, research follows approaches from text retrieval evaluation without questioning whether the achievements in text retrieval performance assessment can be transferred to multimedia evaluation and especially to image annotation evaluation. The hypothesis of this work is that the evaluation of image annotation approaches poses its own demands on test methodologies and evaluation design. These demands will be analysed and experimentally proven.

1.2. Research objectives

The research hypothesis on the fulfilment of special demands in performance assessment of visual annotation approaches can be subdivided into the following research questions:

- **Q1: How is image annotation evaluation performed today?**

System-based evaluation in IR follows the *Cranfield paradigm*. A number of systems are evaluated on a defined test collection that includes query topics, relevance assessed by humans, and test documents. Each answer to a query topic is evaluated by a performance measure and the scores are averaged over all topics. The averaged score is then used to compare systems. While image annotation can be regarded as retrieval task, it poses a

different setup, as each image depicts several topics simultaneously. This research question tackles where we stand in image annotation evaluation and what current limitations are.

- **Q2: Which specific needs should be addressed in the evaluation of image annotation?**

The adequate inclusion of user needs into the evaluation setup is one key issue in system-based evaluation. This question addresses the important point of which user needs should be considered for the evaluation of visual annotation approaches.

- **Q3: What does a user model for concept-based image retrieval look like?**

Performance measures in IR are proposed and applied in dependence on a user model for the retrieval case. This helps to relate the evaluation setup to the user's information need. The question tackles how to define a user model for the retrieval of annotated images and reviews related work.

- **Q4: How can the effort in relevance assessment be reduced?**

Relevance assessment is a key issue in system-based IR evaluation and constitutes a tiring and time-consuming process. This thesis aims at finding cheaper alternatives for relevance assessment that reduce time and effort necessary from experts.

- **Q5: Is it possible and beneficial to include semantics in the evaluation process?**

This question tackles one of the central ideas of this thesis. Images depict several visual topics at the same time. The set of predicted concepts is related to the set of known entities of the image in performance assessment. Current evaluation setups consider each prediction as correct or wrong when judging performance. However, concepts in both sets are semantically related to each other. Pairs of concepts follow different degrees of semantic relatedness from unrelated to perfect match. This thesis investigates how this relation can be used in a fine-grained performance assessment.

- **Q6: Which information source provides the best estimate on semantic relatedness according to desired evaluation characteristics and human judgement?**

This question is closely related to research question Q5. It addresses how to estimate semantic relatedness among visual topics for performance assessment. This work investigates a variety of semantic relatedness metrics and points out limitations of state-of-the-art benchmarks on semantic relatedness estimation for visual topics.

These six research questions constitute the starting point of the thesis. Summarising, the thesis aims at developing new evaluation methodologies for image annotation with a special focus on the inclusion of semantic knowledge in the assessment process while respecting user needs.

1.3. Contributions

The achievements of this thesis fall into three categories. They are related to user requirements and intentions, performance measures for multi-label evaluation, and visual test collections. The major contributions are summarised in the following:

- **Determination of user requirements on photo collection organisation**

The system-based performance assessment of visual annotation approaches should be tightly linked to user requirements in photo indexing. This work aims at taking a step forward in deriving user requirements on organising and tagging photos assessed on a general image collection. The outcomes serve as condition for building visual test collections.

- **Definition of a novel user model for concept-based image retrieval**

User models are a means of including user intentions in system-based performance metrics. Based on a survey, user requirements are collected and comprised in a user model on concept-based image retrieval. This user model determines the basis for the proposal and analysis of performance metrics.

- **Comprehensive review and analysis of performance measures in IR**

This thesis comprises a comprehensive review of IR performance measures in a unified notation. The measures are analysed and categorised according to four dimensions: the measurement direction, the prediction format, the relevance format, and the score prediction dimension.

- **Definition and analysis of novel evaluation measures for multi-label evaluation**

Current performance metrics suffer from fundamental limitations when applied to multi-label image annotation evaluation. These limitations are identified based on the user model for concept-based image retrieval. Three novel performance measures tackle the user requirements by considering the set of predicted concepts and by integrating costs based on semantic relatedness among concepts for misclassifications. The newly proposed measures are related to traditional evaluation measures in comprehensive studies and were successfully included in a real-world image benchmarking campaign. Overall, performance metrics for the different evaluation dimensions are recommended.

- **Low-cost relevance assessment for image annotation**

Relevance assessment is a tedious and time-consuming task in which human experts judge images according to the presence of visual concepts. However, in the assessment of general photos, no special knowledge is needed; this offers great flexibility in how relevance can be judged. Results on the reliability of crowdsourced annotations and on the exploitation of user-generated tags for automated construction of training sets offer a significant chance of simplifying the costly relevance assessment process.

- **Novel image test collection for annotation approaches**

One of the major contributions of this thesis is a fully assessed test collection for image annotation evaluation based on user requirements for photo organisation, comprising 99 visual concepts in 18,000 photos and an ontology.¹ The collection includes affective terms, quality attributes, scene and object descriptions, and representational characteristics of images, and therefore provides a holistic view on images while reducing assessment bias due to the decoupled process of collecting images and defining visual concepts. It was developed and employed in the ImageCLEF benchmarking cycles from 2009 to 2011 and is available to the research community free of charge and without copyright restrictions. This guarantees a large distribution and comprises a highly valuable resource for next-generation image annotation approaches.

- **Achievements of the ImageCLEF Photo Annotation task**

This thesis summarises the experiences, achievements, and outcomes of the organisation of the ImageCLEF Photo Annotation task from 2009 to 2011. The author was responsible for organising the annotation task four times, which includes the definition of the task, the acquisition of the test collection including relevance assessment, the evaluation, and

¹The test collection can be found here: http://www.idmt.fraunhofer.de/de/projekte/abgeschlossene_projekte/photo_annotation.html, last accessed 20.02.2012.

the presentation of results at the workshop. This work incorporates the 180 submissions from three ImageCLEF cycles which constitute a representative selection of the ability of state-of-the-art image indexing systems in 2009 and 2010. The proposed novel performance metrics were successfully adopted as one of the official evaluation measures in 2009, 2010, and 2011.

- **New dataset for semantic relatedness estimation**

Semantic similarity and relatedness estimation has gained a lot of attention within different research fields and with different applications in mind. While datasets for semantic similarity and relatedness estimation, including human judgements, exist for general terms, no semantic relatedness dataset for visual concepts is available. The author proposes a novel dataset to gather semantic relatedness among visual concepts. This resource can be adopted for the benchmarking of semantic relatedness metrics in comparison to human judgement for visual applications.

1.4. Outline of the thesis

The thesis is structured into three main parts. The first introduces the topic of multi-label image annotation and evaluation, and gives the context and background for this work, including a review of state-of-the-art literature. The second part details the contributions on test collections for image annotation: the definition of visual concepts according to user needs for image indexing and retrieval, and the assessment of relevance adopting crowdsourcing approaches. The third part contains evaluation methodologies for image annotation evaluation, including a thorough analysis and evaluation of the proposed performance measures. The test collection and performance measures are adopted in the ImageCLEF benchmark in a real-world scenario for image annotation evaluation. Finally, the contributions and limitations of this work are summarised and thoughts on future work are outlined. The parts are structured into several individual chapters as follows:

Part I: Context and related work

- **Chapter 2** provides an introduction to multi-label image annotation, including the overall process, its basic components, and a review of related work.
- **Chapter 3** details evaluation methodologies in IR and VIR. In particular, it introduces the highly relevant concepts of test collections including documents, topics, and relevance assessments, and summarises IR performance measures.
- **Chapter 4** analyses related work on visual test collections and performance measures in image annotation evaluation. In particular, related work on visual concepts, relevance assessment adopting crowdsourcing, an analysis of common performance measures in image annotation evaluation, performance measures considering semantics, semantic relatedness measures, and meta-evaluation are outlined and open issues are discussed.

Part II: Test collections for multi-label image annotation evaluation

- **Chapter 5** answers which kind of visual concepts should be used to evaluate image annotation approaches based on the results of a user study on personal photo management. It details the requirements of users on image indexing and compares them to the results of other user studies. The findings are used to define a visual concept lexicon and relate the concepts in the form of an ontology.

- **Chapter 6** deals with the costly process of relevance assessment. I analyse whether the ground truth acquisition in image annotation can be outsourced and investigate the reliability of crowdsourced assessments. Effects on system ranking in benchmark-like evaluation and inter-annotator agreement among experts and non-experts are assessed.

Part III: Novel evaluation methodologies for multi-label annotation evaluation

- **Chapter 7** presents the user model on concept-based image search and introduces a novel multi-label evaluation measure that considers fine-grained costs in the performance assessment. In an extensive analysis, the measure is assessed concerning its behaviour on ranking characteristics, random numbers, and over-annotation. It is then compared to standard IR performance measures.
- **Chapter 8** discusses the results of a case study on the ranking and stability characteristics of the novel performance measure when including different sources of semantic relatedness measurement. A new benchmark for semantic relatedness estimation of visual concepts is proposed, and different relatedness measures are assessed and compared to human judgement.
- **Chapter 9** proposes a novel performance measure for ranked predictions that incorporates semantic relatedness estimation. Its characteristics are demonstrated on ranking and stability experiments and it is related to the user model on concept-based search. Besides, the user model is verified in a small user study.
- **Chapter 10** gives a summary on organising, evaluating, and analysing three cycles of the ImageCLEF Photo Annotation task. Research results of this thesis on test collections and performance measures have been incorporated into the benchmark design of the different cycles, and contributions are analysed using the submissions of real participants.

1.5. Publications

Several publications have resulted from the work at this thesis. In the following, the most important ones are listed structured by the chapters they belong to. A full list of related publications is given in the bibliography in appendix A.3. In some publications, the work is described in an earlier stage of development than in this thesis.

Chapter 5: Topic definition: Which concepts do users prefer to organise photo sets?

- Stefanie Nowak and Peter Dunker. A Consumer Photo Tagging Ontology: Concepts and Annotations. In Proceedings of THESEUS/ImageCLEF Pre-Workshop 2009, Co-located with the Cross-Language Evaluation Forum (CLEF) Workshop and 13th European Conference on Digital Libraries ECDL, Corfu, Greece, 2009.

Chapter 6: Relevance assessment: Concept subjectivity and inter-annotator agreement

- Stefanie Nowak and Stefan Ruger. How Reliable are Annotations via Crowdsourcing: a Study about Inter-annotator Agreement for Multi-label Image Annotation. In Proceedings of the ACM international conference on Multimedia information retrieval (MIR), 2010.
- Hanna Lukashevich, Stefanie Nowak, and Peter Dunker. Using One-Class SVM Outliers Detection for Verification of Collaboratively Tagged Image Training Sets. In Proceedings of IEEE International Conference on Multimedia and Expo (ICME), 2009.

Chapter 7: A fine-grained evaluation measure for multi-label annotation evaluation

- Stefanie Nowak and Hanna Lukashevich. Multilabel Classification Evaluation using Ontology Information. In Proceedings of the 1st Workshop on Inductive Reasoning and Machine Learning on the Semantic Web (IRMLeS), co-located with the 6th Annual European Semantic Web Conference (ESWC), Heraklion, Greece, 2009.
- Stefanie Nowak, Hanna Lukashevich, Peter Dunker, and Stefan R uger. Performance Measures for Multilabel Evaluation: a Case Study in the Area of Image Classification. In Proceedings of the ACM international conference on Multimedia information retrieval (MIR), pages 35–44, 2010.

Chapter 8: The effect of semantic relatedness measures on annotation evaluation

- Stefanie Nowak, Ainhoa Llorente, Enrico Motta, and Stefan R uger. The Effect of Semantic Relatedness Measures on Multi-label Classification Evaluation. In Proceedings of ACM International Conference on Image and Video Retrieval (CIVR), July 2010.

Chapter 10: Image annotation evaluation in ImageCLEF

- Stefanie Nowak, Allan Hanbury, and Thomas Deselaers. Object and Concept Recognition for Image Retrieval. In H. M uller, P. Clough, T. Deselaers, and B. Caputo (Eds.), ImageCLEF - Experimental Evaluation of Visual Information Retrieval, The Information Retrieval Series, Chapter 11, Springer, 2010.
- Michael Grubinger, Stefanie Nowak, and Paul Clough. Data Sets created in ImageCLEF. In H. M uller, P. Clough, T. Deselaers, and B. Caputo (Eds.), ImageCLEF - Experimental Evaluation of Visual Information Retrieval, The Information Retrieval Series, Chapter 2, Springer, 2010.
- Stefanie Nowak and Mark Huiskes. New Strategies for Image Annotation: Overview of the Photo Annotation Task at ImageCLEF 2010. In Notebook Papers, CLEF 2010 LABs and Workshop, 22-23 September, Padua Italy, 2010
- Stefanie Nowak. Overview of the Photo Annotation Task in ImageCLEF@ICPR. In D. Unay, S. Aksoy, and Z. Cataltepe (Eds.), Proceedings of the ICPR 2010 Contests, LNCS, 2010.
- Stefanie Nowak and Peter Dunker. Overview of the CLEF 2009 Large-Scale Visual Concept Detection and Annotation Task. In C. Peters, T. Tsirikika, H. M uller, J. Kalpathy-Cramer, J.F.G. Jones, J. Gonzalo, and B. Caputo, editors, Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross-Language Evaluation Forum (CLEF 2009), Revised Selected Papers, LNCS, Corfu, Greece, 2010.
- Stefanie Nowak, Peter Dunker, and Ronny Paduschek. THESEUS Meets ImageCLEF: Combining Evaluation Strategies for a New Visual Concept Detection Task 2009. In Carol Peters, Danilo Giampiccol, Nicola Ferro, Vivien Petras, Julio Gonzalo, Anselmo Penas, Thomas Deselaers, Thomas Mandl, Gareth Jones, and Mikko Kurimo, editors, Evaluating Systems for Multilingual and Multimodal Information Access - 9th Workshop of the Cross-Language Evaluation Forum, LNCS, Aarhus, Denmark, September 2008 (printed in 2009).

Part I.

Context and related work

Outline

This part of the thesis introduces the topic of multi-label image annotation and evaluation and thus establishes the basis for the contributions in this work. Evaluation methodologies for performance measurement in image annotation can only be adequately determined and discussed if the context on image annotation approaches and their most important components are clearly detailed. First, an introduction to multi-label image annotation, including the overall process, its basic components, and a review of related work is given. Then, evaluation methodologies in VIR are outlined in order to set the stage for the definition of methodologies in multi-label image annotation evaluation. Especially, the highly relevant concepts of test collections, including documents, topics, and relevance assessments, as well as performance measures are discussed. Related work on visual test collections and performance measures in image annotation evaluation is described. These works deal with methodologies from text retrieval evaluation if no work is available from within the VIR field. The reviewed literature on image annotation serves as basis to determine and analyse the performance measures which are applied in the evaluation of recent image annotation systems. In particular, related work on visual concepts, low-cost relevance assessment, performance measures in image annotation, performance measures considering semantics, semantic relatedness measures, and meta-evaluation are outlined and discussed.

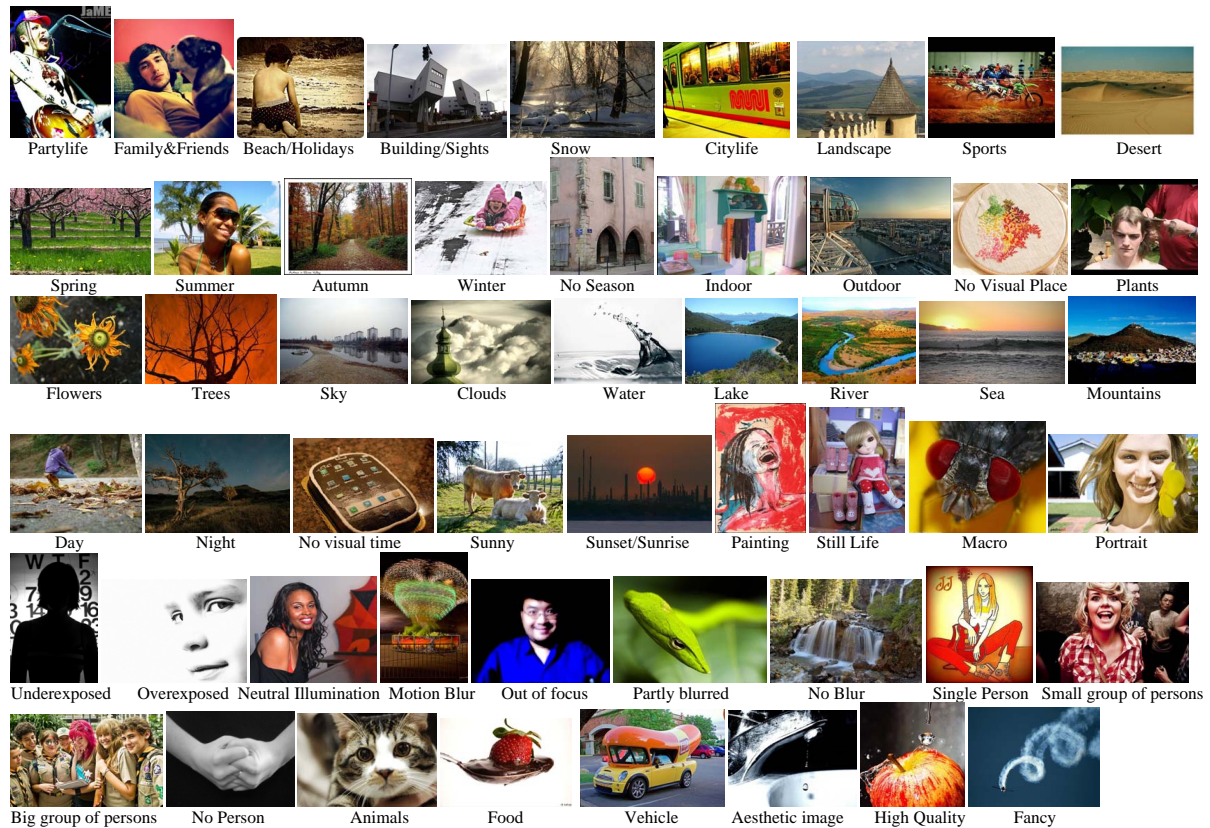


Figure 2.1.: Visual impression of 53 visual concepts that can be utilised for image annotation.

sketches to search for similar images. Please note that this definition of CBIR is stricter than the often cited definition in Datta et al. (2008) where CBIR

is any technology that in principle helps organize digital picture archives by their visual content. By this definition, anything ranging from an image similarity function to a robust image annotation engine falls under the purview of CBIR.

While annotation approaches exist that solely consider textual resources such as captions or surrounding text of images on Web pages, image annotation mostly inherits a content-based component which allows for automated annotation without the need for metadata or other textual resources. Different approaches to image annotation will be highlighted in Section 2.4.

Image annotation is one component of the *concept-based image retrieval* process. Concept-based image retrieval refers to the search in image databases by means of semantic concepts that were automatically detected. In former days, concept-based retrieval was the retrieval of concepts from image descriptions or metadata, and stood in opposition to CBIR which analyses the visual characteristics of images on a low-level basis (see Enser (2000)). In contrast, nowadays concept-based retrieval is understood as interplay of both approaches with the focus on extracting semantic entities of the data. Historically, text-based image retrieval began in the late 1970ies. Images were manually annotated with keywords and the retrieval was performed text-based by using database management systems (see Rui et al. (1999)). The labour intensive work of annotating images manually led to approaches which retrieved images based on their visual content in the early 1980ies. This approach denotes the classic CBIR scenario. Retrieval was performed using an example image or a sketch, and the retrieval engine returned similar images with regard to low level qualities, such as colour or texture. In the following years, research concentrated on approaches that combined keywords from text and low-level visual content for retrieval (e.g. Srihari (1995)), or

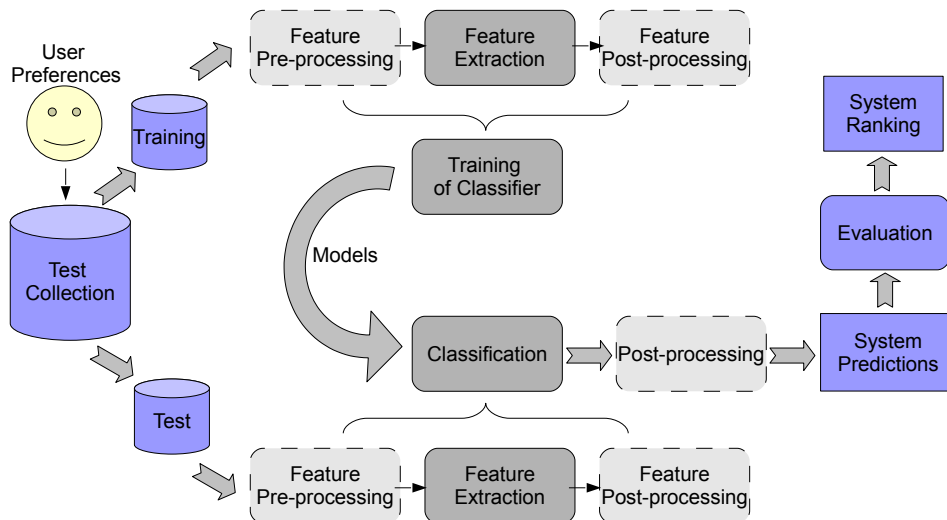


Figure 2.2.: Components of the basic annotation process in context with processing steps of performance evaluation.

approaches that were capable of returning similar images on a higher semantic basis (see e.g., Tieu and Viola (2000) or Weber et al. (2000)). Currently, approaches that are able to automatically index images with high-level concepts based on the visual content are widely explored (see Datta et al. (2008)).

In evaluation, image annotation is often considered by itself rather than in context of the whole retrieval process. This is done in order to judge the quality of the image annotation component without interference of other components in the retrieval process. The following chapters will focus extensively on methodologies for image annotation evaluation. In this chapter, I would like to concentrate solely on the image annotation process. Literature on visual concept-based retrieval, which includes further steps such as a concept selection step and a detector score fusion, can be found in R uger (2009) as well as in Snoek and Worring (2009).

2.2. The basic image annotation and evaluation pipeline

This section briefly highlights the general components of an automated annotation process in connection to the system-based evaluation setup. This consideration focuses on the example of image annotation, but is also valid for media annotation in general. See Figure 2.2 for an illustration of the different components of the system.

The input to an annotation system is a test collection which was designed with specific user needs in mind. The test collection contains the visual documents, a list of concepts to be detected in the media items, and relevance judgements for at least a part of the collection. Chapter 3.2 will provide an extensive discussion of test collections. Usually, the test collection is divided into two or three parts: a training set, an optional optimisation or validation set, and a test set. The training set is used for the training of the classifier which may be tweaked on the validation set. The test set is considered in the test of the classifier which evaluates its annotation quality. The classifier is only optimised on the training (and validation) set in order to guarantee its applicability to unknown document collections. An optimisation on the whole test collection would lead to an over-fitting of the classifier and to low generalisation capabilities.

The output of an annotation system consists of predictions about the presence of concepts for each media item. These predictions may be likelihoods, thresholded confidences, or binary

predictions, depending on the chosen classifier and the system setup. These predictions are assessed by a performance measure in the evaluation process. The resulting score can be used to compare the system's annotation performance to those of other annotation systems which were judged on the same test collection.

In essence, the annotation system consists of two processing steps. The first is called feature extraction. This is the process of extracting visual signatures of the media items or of generating features from metadata such as exchangeable image file format (EXIF) tags or user-generated tags. These features further represent the characteristics of the item in a condensed form. Feature generation may be accomplished by a feature pre-processing and a feature post-processing step. The features are employed in the second processing step, the classifier training and classification step. During training, the classifier uses the features and the relevance judgements to learn a classification model. In the test, the same features are extracted from the test media items and the learned model is utilised to predict the concepts.

Finally, some annotation systems additionally consider external knowledge. Often, knowledge bases such as the Web, ontologies, or other taxonomies are applied to determine relations between concepts. These may be exploited in a post-processing step after the classification. In contrast to metadata, external knowledge is not directly associated with the particular media item.

In the following, the two main components in image annotation, descriptors and classifiers, are described in detail and information fusion approaches are highlighted.

2.2.1. Descriptors

Descriptors determine which image characteristics are extracted and how they are compactly represented for further processing steps. While the term "descriptor" refers to the specification of the mathematical calculation of image characteristics, the terms "features" and "visual signatures" are used to describe the actual computed characteristics of an image.

The applicability of descriptors is limited due to the so-called *semantic gap*. Smeulders et al. (2000) define the semantic gap as

the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation.

Furthermore, descriptors aim to solve the *sensory gap*, which denotes the difference between the digital representation of a concept and its real-world counterpart. Digital representations may suffer from lightning variations, viewpoint changes, or occlusions. Therefore, descriptors aim at a representation which is invariant to lightning changes or transformations (translation, rotation, scaling) while retaining their discriminative power.

In image annotation, descriptors consider textual information (i.e., metadata, captions, or tags), the visual content of images, or a combination of both. Visual descriptors either consider the images as a whole (global features) or take only a region of the image into account (local features). Common global descriptors are based on colour or texture, such as colour histograms (in different colour spaces), colour layout or dominant colours, and edge histograms, or texture descriptors based, for instance, on wavelets or the Discrete Cosine Transform. Further, shape is considered as an important characteristic which is determined either contour-based or region-based. Often, shape-based descriptors require a segmentation step as pre-processing. As a reliable segmentation for unconstrained image annotation is an unsolved problem, research considers weak segmentation methods and local descriptors which are computed on a pixel or a small block of pixels including its neighbourhood. A variety of visual descriptors has been standardised by the Moving Picture Experts Group (MPEG) in MPEG-7 (see Manjunath et al. (2002)).

The most popular local descriptor is the Scale-Invariant Feature Transform (SIFT) proposed by Lowe (2004). This local invariant descriptor has proven to be robust in object detection tasks

and is therefore widely adopted. Research extended the SIFT descriptors, which were originally proposed to work on grey values, to colour descriptors. Bosch et al. (2008) propose the HSV-SIFT descriptor. In this descriptor, the SIFT features are calculated for each hue, saturation, and value (HSV) component. Abdel-Hakim and Farag (2006) investigate the C-SIFT descriptor, which uses the colour invariant, to derive the SIFT features. The colour invariant normalises the opponent colour space by deleting its intensity information and therefore makes it scale invariant with respect to light intensity. An extensive review on colour SIFT descriptors can be found in van de Sande et al. (2010).

In contrast to global descriptors, local image descriptors consider different regions of interest in varying sizes. Some local descriptors use small square image regions while others consider large portions of the image. The positions from which local features are extracted vary widely dependent on the adopted sampling strategy. Local features can, for example, be extracted from sparse interest points or from a dense grid. The local features are often summarised in a bag-of-visual-words (BoW) descriptor. Many variations of BoW approaches have been proposed. The most common approach consists of clustering a set of representative local features using k -means into 500-4000 cluster prototypes. These cluster prototypes are denoted as visual words and form a visual codebook. Then, each image is represented by a histogram, counting how many of its local features belong into which of the cluster's quantisation. Van Gemert et al. (2010) perform an extensive study on the assignment of features to visual words. This codebook model drastically reduces the dimension of the local features into a compact histogram. Jiang et al. (2010) explore the influence of different representation schemes for BoW on annotation performance, such as vocabulary size, weighting scheme, or feature selection.

The global and local descriptors introduced so far belong to the group of low-level and mid-level features, respectively. They represent global and local characteristics of the image, but have no direct semantic meaning concerning the properties or entities that are pictured. However, there are approaches that attempt a direct semantic modelling of properties and entities in the images. Oliva et al. (2001) define five perceptual qualities: the naturalness, the openness, the roughness, the expansion, and the ruggedness to model the spatial structure of a scene. Further descriptors aim at the detection of special quality characteristics in images, such as blur or aesthetics (see Ke et al. (2006) and Datta et al. (2006)). Extensive reviews on visual descriptors may be found in Smeulders et al. (2000) and Datta et al. (2008).

Most image annotation approaches solely consider visual content due to the lack of a textual embedding of images. With the advance of social network sites such as Flickr¹, images are accompanied by user-generated tags at no costs. Although these tags often contain random noise, and are usually imprecise, personalised, and limited, they can be exploited as features for automated image indexing. Hwang and Grauman (2010) analyse the tag lists of images in order to determine implicit tag features, such as the prominence of an object derived from the position in the tag list, the spatial proximity estimated through the sequence in the tag list, and the layout of objects based on assumptions from gaze theory. Several approaches consider a binary presence/absence vector for the most frequent Flickr tags in the test collection, as in Verbeek et al. (2010). Other work focuses on the detection of tag ambiguity (Weinberger et al. (2008)) or on the analysis of user-generated tags and tag sets (Bischoff et al. (2010), Huang and Zhou (2010)).

2.2.2. Classifier

The problem of automated image annotation is formulated as a machine learning problem. Systems adopt unsupervised or supervised machine learning approaches in order to classify unknown

¹www.flickr.com

images and to label them with appropriate concepts based on the concept information of the training set and on the extracted features. Unsupervised approaches include clustering methods, such as k -means clustering, or topic models, such as the Latent Dirichlet Allocation (LDA) (although there exist supervised variants of topic models (Blei and McAuliffe (2007))). Supervised classifiers are often grouped into generative, discriminative, or model-free approaches. Generative probabilistic models estimate the distribution of features for each class and use this to predict which class is most likely for a certain observation. Generative models include Gaussian Mixture Models (GMMs), Markov Random Fields (MRFs), or Hidden Markov Models (HMMs). Discriminative approaches directly model the classes. The distance to the class boundaries can be used to model posterior probabilities for the classes. The Support Vector Machine (SVM) is the most commonly adopted discriminative model for multimedia annotation and represents the default choice. Further discriminative models include log-linear models, logistic regression, decision trees and random forests, Conditional Random Fields (CRFs), and neural networks. Prominent model-free approaches include variations of the nearest-neighbour approach (NN). Another option consists of combining or blending generative and discriminative approaches.

In multi-label classification, several techniques to model the simultaneous membership of one image to different classes are adopted. The desired outcome of an image annotation system is a bipartition of the concepts into a set of relevant labels and a set of irrelevant labels for each unknown image (binary output), or a ranking of the concepts according to their probability of occurrence. According to Tsoumakas et al. (2010), multi-label classification approaches can be distinguished into problem transformation approaches and algorithm adaptation methods. Problem transformation approaches include binary relevance methods which learn one classifier per concept, transformation methods which transform the multi-label dataset into a single-label multi-class dataset and learn a single multi-class classifier, or selection strategies that choose one label out of several ones. Algorithm adaptation methods extend traditional classification approaches in order to directly handle multi-label data. If a concept hierarchy is available, hierarchical multi-label classifiers can be trained instead of flat multi-label classifiers. Hierarchical classification approaches include the training of a binary classifier on each non-root node which considers the training examples belonging to the subtree, the training of a multi-label classifier on each non-leaf and non-root node, or predictive clustering trees and decision trees.

Image annotation poses heavy demands on the classifier. Usually, systems have to deal with an imbalance between the number of positive and negative examples per class and with a limited number of training examples in general; they have to face the curse of dimensionality problem and to avoid over-fitting. The curse of dimensionality refers to the problem that the dimension of the concatenated features is too large in comparison to the number of training examples (see Jain et al. (2000)). Over-fitting typically indicates that the classifier is over-optimised on the training set. Both issues result in a poor generalisation to unknown test examples.

2.2.3. Fusion techniques and post-processing

Image annotation approaches mostly consider a set of features based on the assumption that different descriptors extract complementary information. This includes several features which were extracted from the same modality, as well as multi-modal features. In image annotation, features from the visual and the textual modality can be considered. Fusion techniques aim at the combination of all information acquired in the annotation process in order to come to the final classification decision. Fusion approaches may be incorporated on the feature level or on the classifier level; this is also referred to as *early fusion* and *late fusion* approaches.

Early fusion combines different features in the feature space before the classification is performed. Then, one classifier is trained that may directly operate on multi-modal data if the

features were extracted from different modalities. Early fusion approaches require only one learning phase as only one classifier is trained, but have to deal with problems arising through different numerical scales of the features and the size of the concatenated feature vectors. Classification with increased feature dimension might run into the curse of dimensionality problem if the number of training examples is not sufficient. To reduce the size of the concatenated features, feature transformation methods are often applied. Common techniques include the Principal Component Analysis (PCA) (see Jolliffe (2002)) and the Linear Discriminant Analysis (LDA*) (see Fukunaga (1990)). The latter of these techniques requires knowledge about the associated concepts, as it maximizes the ratio of between-class variance to within-class variance. Both approaches allow discarding feature dimensions as they transform the feature matrix so that it contains the dimension with largest variances on top. Other approaches consider a feature selection strategy to determine the feature dimensions with greatest Information Content (IC).

In late fusion approaches, the output of a set of classifiers is combined. This includes approaches that train a classifier for different subsets of features (i.e., for each modality or for each descriptor), techniques that consider different classifiers (i.e., simple classifiers for simple concepts and advanced classifiers for difficult concepts), and classifiers that consider different subsets of the training data (i.e., bagging and boosting approaches). Late fusion approaches require a learning phase for each classifier, but have the general advantage that the accuracy of the classification result can be increased if complementary information is used. The classifier outputs are combined based on the concept rankings, the concept decisions, or the confidence scores. Combination methods either consider a supervised approach in which all classifier outputs are used as input for a final classifier, or unsupervised approaches that use combination rules such as the majority vote, maximum, minimum, or geometric mean to combine classifier outputs.

Depeursinge and Müller (2010) analyse fusion techniques for the fusion of textual and visual information that have been applied by participants of the ImageCLEF benchmark between 2004 and 2009. They conclude that most participants utilised a late fusion approach, possibly in order to optimise textual and visual retrieval systems separately and to circumvent the curse of dimensionality problem. The majority of approaches consider a linear combination of confidences, often with arbitrarily chosen thresholds. Snoek and Worring (2009) review fusion approaches for video annotation, while Tsai and Hung (2008) detail the use of multiple classifiers in image annotation literature.

Beside the early and late fusion techniques, there also exist approaches which model the relations between different concepts in a post-processing step and use this information to enforce or to lessen confidence scores for particular concepts. Applied approaches include co-occurrences between concepts derived from the training set and external knowledge in the form of ontologies.

2.3. Explicit semantics in Visual Information Retrieval

The use of explicit semantics in image annotation was adopted in order to tackle the semantic gap problem. Models of the annotation domain are built and used in the verification of the annotation decisions. Often, these “real-world models” are encoded in the form of ontologies. According to Gruber (1995), “an ontology is an explicit specification of a conceptualization” in which the conceptualization refers to an abstract, simplified view on a part of the (imagined) world. Gruber (2009) later extends this definition by determining:

An ontology defines (specifies) the concepts, relationships, and other distinctions that are relevant for modeling a domain. The specification takes the form of the definitions of representational vocabulary (classes, relations, and so forth), which provide meanings for the vocabulary and formal constraints on its coherent use.

This definition clearly states that the semantics of an ontology do not necessarily capture the complete semantics of a described object, but that the semantics of an ontology define the coherent usage of the object's semantics in a special context. Ontologies therefore allow the interoperability between applications by exchanging machine-readable information in a unified presentation. Dependent on the broadness of the represented knowledge area, Guarino (1998) classified ontologies into core, domain, task, and application ontologies. Further reading on Semantic Web technologies, including ontology description languages such as the Web Ontology Language (OWL), ontology query languages, or ontology engineering, can be found in Gomez-Perez et al. (2003), Hitzler et al. (2009), or Breitman et al. (2007).

Related approaches model semantics in the form of *vocabularies*, *taxonomies* or *thesauri*. These knowledge representations vary according to the degree of semantics that can be expressed in the model of the real-world domain. Vocabularies contain an unstructured enumeration of concepts which are not explicitly related to each other. Taxonomies structure vocabularies in a hierarchy which allows grouping several concepts or to model father-son relationships. Thesauri may use more extensive relationships and associative rules than taxonomies, which are fixed in the corresponding ISO standard (ISO2788 (1986)). In contrast to thesauri and taxonomies, ontologies are able to express arbitrary relationships, such as restrictions in cardinalities, restrictions in values, disjointness of concepts, datatype restrictions, or rules; therefore they have a higher expressive power. Several core ontologies with linguistic foundation such as WordNet (Fellbaum (1998)), ConceptNet (Liu and Singh (2004)), or Cyc (Lenat and Guha (1989)) represent commonsense knowledge in a machine-readable format.

2.4. Related work on multi-label image annotation

Related work on multi-label image annotation can be categorised in different ways. Research could be divided into what kind of features are extracted, which classifier is applied, which multi-label classification strategy is adopted, on which test collection the performance has been evaluated, or which domain is targeted. As the literature on image annotation and concept-based retrieval from the 1980ies on is extensive, this section exemplarily summarises the recent streams for automated annotation, focusing especially on the information sources used in annotation. Interactive approaches (i.e., employing relevance feedback), categorisation approaches, similarity retrieval, or works explicitly focused on video annotation, as well as a complete discussion on image annotation history are omitted. Image annotation imposes no restrictions on what kinds of concepts are annotated. It therefore subsumes research fields as natural scene recognition (Boutell et al. (2004)), object detection (Fergus et al. (2005)), or medical image annotation (Tommasi and Deselaers (2010)).

The majority of work on image annotation considers content-based approaches. Mori et al. (1999) extract colour and edge features on rectangular parts of the images and perform a vector quantisation to cluster these features. Each sub-image is associated with all concepts of the image. The likelihoods for each cluster are based on the frequency of labels and are used to determine the most probable concepts for a test image. Duygulu et al. (2002) extend this work in order to assign concepts to image regions. The authors examine the problem of image annotation as a machine translation process in which visual features are translated into concepts with a translation model; this is analogous to translating terms of one language into another with the help of a lexicon. Images are segmented with the normalized cuts algorithm and the resulting regions are clustered into 500 blobs based on various low-level features. The lexicon is learned on the basis of correspondence between concepts and blobs in the training set using the Expectation-Maximization (EM) algorithm. Fan et al. (2004) perform the image annotation process by first detecting salient objects based on a mean-shift segmentation and then employing a SVM classifier

using low-level regional colour, shape, and texture features. After determining the salient objects, low-level features are extracted of the object area and a finite mixture model is used to annotate the image with scenes. Feng et al. (2004) use multiple Bernoulli relevance models to learn associations between concepts and images based on colour and texture features extracted from rectangular tiles. Yavlinsky et al. (2005) perform image annotation based on nonparametric density estimation. They adopt simple global colour and texture features and the Earth Mover's Distance (EMD). Carneiro and Vasconcelos (2005) as well as Carneiro et al. (2007) propose a supervised multi-class labelling approach which learns a semantic class density with a hierarchical estimation procedure, consisting of a Gaussian mixture per image and an extended EM algorithm.

Several extensions of the LDA model (Blei et al. (2003)) have been proposed in order to better capture joint distributions of image and caption words in multimedia documents. Blei and Jordan (2003) present correspondence Latent Dirichlet Allocation (corr-LDA), which models the conditional distribution of the annotation given either the image regions or the concepts. For each caption word, a hidden topic variable is directly shared with a randomly selected image region. Barnard et al. (2003) propose multi-modal Latent Dirichlet Allocation (mm-LDA). This approach also models the joint distribution of image regions and concepts, but shares the mean topic proportion variable between both modalities. Putthividhy et al. (2010) propose the topic-regression multi-modal Latent Dirichlet Allocation (tr-mmLDA) model, another multi-modal extension of the LDA, which learns the joint correlations between image features and annotation texts by two topic models and a latent variable regression approach. In contrast to mm-LDA, it does not assume that image and caption words are exchangeable and therefore exhibits a higher prediction quality.

Recent research investigates label propagation approaches and finds them to be competitive in annotation performance while reducing computation times. Makadia et al. (2008, 2010) demonstrate that a simple k -NN approach for label propagation with global colour features is able to outperform state-of-the-art algorithms. These results can be improved by using a weighted NN model that predicts labels based on a weighted combination of their presence and absence among neighbours, as presented by Guillaumin et al. (2009) in their TagProb system. Mei et al. (2008) follow a label propagation method and make use of semantic distances to annotate images. They cluster the annotations of the training set into semantic clusters and determine a semantic distance function for each cluster which maps global visual features in semantic distances. The semantic distance is derived from WordNet. For new test images, the annotations are propagated based on the cluster memberships. Chen et al. (2010) present an efficient graph-based multi-label propagation approach for large-scale multi-label annotation which is based on a hashing scheme for graph construction. The similarity measurement for label propagation uses the Kullback-Leibler divergence and local visual features such as colour moments, wavelet texture, and edge histogram. The algorithm outperforms several state-of-the-art graph-based approaches, as well as a SVM classifier with respect to both accuracy and computation times.

While most work on automated annotation focuses on the increase of prediction accuracy, Li and Wang (2007) propose an approach for real-time tagging of Web images. They adopt a segmentation into regions based on wavelets and k -means algorithm and extract colour and texture features. Novel statistical methods for feature modelling and model matching that are able to deal with discrete distributions are presented. Performance is tested in a cross-collection setting and evaluated using Web images and user judgements.

Research also considers the integration of user-generated tags, such as Flickr tags, into multi-label image annotation. In particular, Ling et al. (2008) determine the relative importance of tags based on their frequency and their correlation in a collection and use them as prior probability on a graph-based method for image annotation. Huang and Zhou (2010) explore possible relationships among Flickr user tags. They define four different relation types: parent-child relations, hypernym-hyponym relations, token-phrase relations, and synonym relations. Their work reveals that the

incorporation of these relations can help improving tag clustering significantly. Motohashi et al. (2010) propose conceptual fuzzy sets for image annotation using textual and visual information. The textual features are based on term frequency - inverse document frequency (tf-idf) weights for each tag per concept. Then, the tf-idf weights are summed up in order to obtain confidence scores for each image. Verbeek et al. (2010) show that the inclusion of Flickr tags as features helps in increasing performance by about 11% MAP in contrast to a visual run. They follow an early fusion approach in which a binary presence/absence vector of the 457 most frequently occurring Flickr tags is concatenated to the visual features. Liu et al. (2009) use the tag sets of images from social networks sites such as Flickr to estimate the relatedness of the tag set to a query concept. The relatedness is estimated using Flickr co-occurrence and Wikipedia, as well as visual similarity based on colour and texture features. Duan et al. (2009) present an annotation system that considers the style of images in order to predict annotations. Images belonging to the same batch are considered to refer to one style. For each style, a separate annotation model is trained, using probabilistic latent semantic analysis (PLSA) based on the grouping information and SIFT features. Xu et al. (2009) use Flickr's Related Tag resource to learn the semantic correlation among visual concepts. They adopt CRFs to model different information sources within one unified framework. Features are derived from the visual content, the textual embedding of images, and the keyword correlations of Flickr.

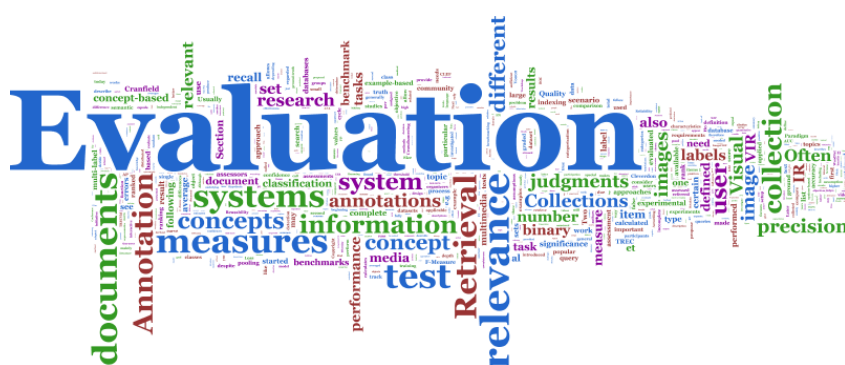
Other research streams deal with automated annotation by adopting ontologies. Jin et al. (2005) adopt the translation model of Duygulu et al. (2002) and prune irrelevant keywords using semantic distances on WordNet. Srikanth et al. (2005) explore a hierarchical classification approach which exploits hierarchical dependencies between concepts derived from WordNet. Each node is associated with one image region (blob) which was generated with a k -means clustering on image regions. Fan et al. (2007, 2008) propose another hierarchical classification approach using a concept ontology and global and local visual features. The concept ontology is constructed based on the visual features of the data and the distance in WordNet between categories. A multiple kernel learning algorithm trains SVM classifiers for the atomic image concepts of the concept ontology. A hierarchical boosting algorithm then incorporates the concept ontology and multi-task learning in order to train hierarchical image classifier models. Datta et al. (2007) propose a lightweight meta-learning framework for image annotation over time. The output of a black-box annotation algorithm is improved by considering available information, such as user feedback and WordNet knowledge, in an efficient re-training scenario.

Further notable works include the non-parametric Continuous-space Relevance Model (Lavrenko et al. (2003)), the cross-media relevance model (Jeon et al. (2003)) and its extended version for multi-modal features (Wang et al. (2009)), PLSA for image annotation (Monay and Gatica-Perez (2004)), the multi-edge graph (Liu et al. (2010)), or the Image-Concept Distribution Model (Su et al. (2010)). Numerous overviews of the state-of-the-art for visual retrieval and annotation have been published, i.e. in Tsai and Hung (2008), Datta et al. (2008), Lew et al. (2006), Smeulders et al. (2000), Bosch et al. (2007), Liu et al. (2007), or Rui et al. (1999).

2.5. Summary

This chapter has provided an extensive insight into multi-label image annotation and its basic components. Recent research was reviewed with a special focus on the variety of information sources that are employed to improve prediction quality. However, the important research area on evaluation methodologies of annotation systems, available test collections, evaluation metrics, and benchmarks was not addressed in this chapter. Due to its importance for this thesis, this topic will be discussed in a separate chapter which provides a detailed discussion on image annotation evaluation, including its basis in IR evaluation.

3. Evaluation of visual Information Retrieval and annotation



This chapter provides an introduction into the field of IR evaluation with a special focus on VIR evaluation and annotation. Its goal is to introduce the relevant terminology and to summarise the main research focus. A thorough review of IR evaluation is given by van Rijsbergen (1979), Voorhees and Harman (2005), Manning et al. (2009), Sanderson (2010) and Harman (2011), while the work of Müller et al. (2010) focuses on VIR evaluation. The chapter is organised as follows: after a short review of the history of experimental evaluation, the chapter deals with reference collections in Section 3.2, including requirements, visual concept definition, relevance assessment, and a summary of multimedia collections. Further, Section 3.3 details the most important evaluation measures, including a brief introduction to significance tests. Finally, the chapter closes with a description of recent benchmarking campaigns in Section 3.4 and a short summary in the ultimate section.

3.1. History of experimental evaluation

Evaluation is the process of determining the performance and benefits of the object of investigation. In IR, evaluation assesses different aspects of IR systems, such as the extent to which user needs are addressed, the degree to which a system correctly solves a task, or processing times. IR evaluation aims to establish a comparison between different systems in a defined scenario.

Two complementary approaches of IR evaluation exist: the system-based evaluation and the user-based evaluation. In both cases, the fulfilment of the user's information needs is evaluated. In the first approach, the system capabilities of ranking and indexing are investigated, while the second approach measures user satisfaction and system-user interaction. Although the latter approach allows measuring user satisfaction more directly, it is expensive and difficult to perform, as detailed in Voorhees (2002). System-based evaluation makes the assumption that the user's information need can be measured by an appropriate ranking of documents in which higher ranks denote a higher relevance of documents. This abstraction allows for a cheaper evaluation scenario

and delivers quantitative data on system behaviour. In the following, the focus is placed on system-based evaluation of IR systems. System-based evaluation is also referred to as *laboratory evaluation*, *experimental evaluation*, *quantitative evaluation*, or *objective evaluation*.

While experimental evaluation in IR began in the early 50ies, the first large-scale laboratory evaluation was performed between 1963 and 1966 with the Cranfield II experiments by Cleverdon et al. (1966), Cleverdon (1967), and the SMART project of Salton (1971). The Cranfield experiment became popular due to the fact that it was also the first experimental evaluation that produced unexpected results. Cleverdon et al. (1966) found out that single-term indices outperform simple concept indices (word phrases) and thesauri (controlled vocabularies) in an indexing experiment with about 1400 documents and 300 queries. This violated the expectations of librarians and experts that believed in sophisticated indexing languages and classification schemes and resulted into a controversy about the experimental setup (see Salton (1992)).

In the following years, the style in which Cleverdon formulated his indexing experiments was adapted for IR evaluation and is today generally referred to as the *Cranfield Evaluation Paradigm* (Voorhees (2002)). The Cranfield paradigm postulates that the evaluation of different systems should be performed on the same set of documents and regarding the same set of information needs, which should be defined before the start of evaluation. The definition of a user model comprises the core of the paradigm containing the information need, real questions, and appropriate metrics that reflect the user. In Cleverdon's time, Precision and Recall (see Section 3.3) were used to evaluate the effectiveness. Cleverdon used domain experts that posed questions and assessed relevance of documents based on topic similarity. While he manually evaluated indexing schemes for aeronautic engineering documents, today, experimental evaluation is mostly performed in an automated manner.

According to Voorhees (2002), the Cranfield paradigm makes three assumptions to allow for quantitative evaluation. First, it is assumed that relevance can be approximated by topical similarity, which also implies that all relevant documents are equally desirable to be retrieved, that the user need is static, and that the relevance of documents is independent from other documents. Second, a single set of judgements for a topic is regarded as representative of the user's information need, and third, the list of relevant documents for one topic has to be complete. Of course, these assumptions do not generally hold; as a consequence, experimental evaluation introduces noise into the ranking process. The noise can be reduced dependent on the number of topics and the choice of evaluation measures. While the evaluation score for a particular system cannot be regarded in isolation, this setup allows for a comparison of systems that were tested with the same test collection. Cleverdon stressed the importance of using a complete set of judgements, but today's IR evaluation experiments mainly make use of incomplete judgements and techniques have been developed for a stable comparison of systems in the absence of a complete set of judgements (see Section 3.2.3). Further, Cleverdon's set of test documents was rather small, while today, the need for a larger collection is proven in order to generate representative results, which makes a complete relevance assessment intractable as shown by Baillie et al. (2008).

The Cranfield paradigm is applied to most IR evaluation initiatives today; it also describes the current experimental setup for VIR and annotation evaluation. The history of VIR evaluation began in the early 1990s, 30 years after the Cranfield experiments. Müller et al. (2001) state that in the beginning, it was common practice to only plot a screenshot of the system output in research papers to allow for a visual inspection of the retrieval results, as for example in Picard and Minka (1995), and Zhao and Grosky (2000). The utilised datasets were mostly collected by the author of the annotation algorithm and contained just a few images with a few annotations; they were generally not available to other research groups. The need for experimental evaluation with a defined test collection that would allow comparing different systems was stressed by the end of the 90ies. The idea that many solutions from IR evaluation could be adapted for VIR despite the

differences between the fields became obvious, and their suitability was systematically investigated as shown by Smith (1998), Narasimhalu et al. (1997), Müller et al. (2001) and Grubinger (2007).

3.2. Test collections for VIR

A test collection consists of three distinct components: the document collection, the topics (which are often referred to as *concepts* in annotation tasks), and the relevance assessments. The relevance assessments are represented as a list of document IDs. The list determines which documents should be retrieved for a certain topic. In the following, requirements on document collections are summarised, the definition of visual concepts is introduced, and possibilities of obtaining relevance assessments are described. The section ends with a summary of test collections for VIR and annotation tasks.

3.2.1. Requirements

The collection of databases and the definition of a ground truth is primarily hampered by the high ambiguity in annotating multimedia documents and the amount of time needed. Requirements on the test collections are generally high and strongly task dependent. The following requirements and specifications for the creation process of image test collections have been defined by Grubinger et al. (2010) and Datta et al. (2008):

1. Representativeness: The collection should be representative of the particular domain in question. It should contain the types of images and annotations that are applicable to state-of-the-art systems.
2. Diversity: The collection should contain a diversity of subject matter. As collections are often domain specific, they should at least represent a variety of documents for this particular area.
3. Evaluation scope: The collection has to fit the evaluation objectives.
4. Collection size: The quantity of images and annotations should be applicable to state-of-the-art systems and should provide the possibility of extracting meaningful query topics to evaluate the state-of-the-art retrieval methods. The collection has to be large enough to allow for significance testing.
5. Collection parameters: The collection shall be capable of being parameterised according to the evaluation needs.
6. Document quality: The quality and resolution of images, videos, or annotations should be applicable to state-of-the-art systems and the evaluation objective.
7. Document metadata: Additional metadata such as image annotations, captions, or user tags should be available because they provide high flexibility in the task design. They allow to benchmark multi-modal systems that do not rely solely on content features, or to compare systems using different information sources on the same collection.
8. Copyright: In multimedia research, copyright restrictions are an important issue. Ideally, benchmark collections are free of copyright restrictions, so that they can be distributed to research groups and used in papers.
9. Reusability: The reusability of test collections beyond their initial evaluation objective is a desirable property.

10. Ground Truth: Beside the documents themselves, the ground truth is arguably the most important part of the test collection. A complete and consistent ground truth that can be distributed together with the documents is ideal.
11. Reliability: The reliability of a test collection describes to what extent differences in system predictions are not due to chance. The design of a test collection is influenced by the documents, the assessors, the queries, and the interaction between these components. With a few sample data, the reliability of a test collection can be estimated prior to its design by applying statistical approaches such as test theory (Bodoff (2008)).

3.2.2. Visual concepts

The evaluation of IR and VIR depends on a user model as described in Section 3.1. In the evaluation setup, a distinction between the information need and the actual query is made. The information need comes in the form of a statement and is generally called *topic*. It often consists of four parts: identifier, title, description, and narrative (Voorhees (2002)). The query itself is the actual formulation of the information need that is given to the system. For visual annotation evaluation, there is often no distinction between the concept and the query, as the general procedure is different from the IR and VIR evaluation scenario. In an annotation task, participants are provided with a training set consisting of the documents, the ground truth, and a list of concepts. Classifiers learn these concepts and later predict them on the test set. For a detailed description, see Chapter 2. Despite this different setup, the visual concepts should also adhere to a user model and describe an information need.

In the beginning of research on image categorisation and annotation, the evaluation was mostly performed on small test collections, which were acquired by the author of the annotation approach and ordered into different semantic classes. Zhao and Grosky (2000) use ten semantic classes (*towers, ancient columns, birds, horses, pyramids, rhinos, sailing scenes, skiing, sphinxes, and sunsets*) for each of which they collected five images. Vailaya et al. (1999) test their classification performance on the disjoint classes *indoor versus outdoor* and *landscape versus city*, while Weber et al. (2000) evaluate performance on the objects *leaves, car, and head*. In the works reported here, no motivation or further information on the reasons to choose these kinds of categories and objects is available. It is also apparent that some approaches focus more on scenery descriptions of images, while others address object detection; further work is interested in objects as well as in holistic image descriptions. Picard and Minka (1995) discuss problems that arise when predicting which labels a user may want to assign to a region of an image. Depending on the viewpoint of the user, the semantics, the culture, or the visual features, they may choose a different label for an image region. The authors therefore propose an extension of their system that allows users to define labels on regions of images. The system predicts similar regions in other images and employs the user input as relevance feedback. They do not provide an answer to the question of which concepts a user may choose, but circumvent the problem with an interactive system.

Research on the type of semantic concepts users may prefer was initiated quite early by Eakins and Graham (1999), Jörgensen (1998), and Greisdorf and O Connor (2002), and theories and user studies were performed to collect empirical data. Hollink et al. (2004) investigate which categories of image descriptions exist and how much users adopt them to formulate image queries. Laine-Hernandez and Westman (2006) concentrate on the description and categorisation of journalistic images. Most of these studies aim at obtaining a holistic picture of how users process and describe images and how search systems could support them. The answers are not restricted to purely visual concepts. Further, these works do not provide the datasets and annotations that are needed for training and test of the annotation approaches.

In around 2006, efforts were made to combine the results of user studies and requirements to create exhaustive lists of visual concepts, and to collect images per concept for training and test purposes. Further, the focus was on *visual concepts*, i.e., concepts that are potentially detectable with content-based indexing methods. Naphade et al. (2006) and Kennedy et al. (2006) propose the Large-Scale Concept Ontology for Multimedia (LSCOM). LSCOM contains 1,000 concepts for describing broadcast news videos that were identified in a series of workshops, and provides annotations for 449 of them. Snoek et al. (2006) define 101 semantic concepts for multimedia annotation on the TRECVID corpus in a lexicon. Loui et al. (2007) construct the Kodak benchmark dataset for the detection of semantic concepts in consumer videos. Its concepts are derived from findings of user studies. In related work, visual concepts are not derived from user studies but from a large linguistic corpus of the English language: WordNet (Fellbaum (1998)). In WordNet, English nouns, verbs, and adjectives are organised into synonym sets which all represent an underlying lexical concept. Deng et al. (2009) construct a database with 10,184 categories that were derived from the approximately 80,000 noun synonym sets of WordNet. Xiao et al. (2010) also used WordNet nouns to describe 899 image scenery categories in about 130,500 associated images.

3.2.3. Relevance assessment

Closely related to the problem of finding adequate visual concepts for the annotation evaluation task is the task of defining which kind of images are relevant for a particular concept, and which are not. In general, relevance determines the attribution of a certain image to a certain concept. Formulated differently, relevance should determine whether an image should be returned to users, if they include the concept in the query formulation. For example, in Text REtrieval Conference (TREC), an entire document was defined as relevant if users included any information from a document in a report on a certain topic, regardless of other documents providing the same information (Voorhees (2002)). Of course, this is a rather pragmatic definition of relevance. Greisdorf (2000) proposes an interdisciplinary view on how relevance can be defined, and Mizzaro (1997) provides a thorough review of literature on relevance. There is a distinction between binary relevance (i.e., a document is either relevant or not) and graded relevance. Cleverdon used a 5-point graded relevance scheme in his experiments, while most of the evaluation benchmarks today (see also Section 3.4) stick to binary relevance. The way in which relevance is assessed also influences the choice of evaluation measures that are applicable. Basically, there are three issues entailed in the assessment of relevance: the *consistency*, the *completeness*, and the *reusability* of judgements.

3.2.3.1. Consistency

Cuadra and Katter (1967) as well as Harter (1996) show that different assessors produce a different set of relevance judgements. The reasons are manifold, but reside to a certain amount in the fact that often, the border between relevant and irrelevant items cannot be drawn exactly. As a result, some decisions are made by the subjective interpretation of the assessor. This problem remains although a definition is provided per concept about what should be regarded as relevant. Further, the opinion about the relevance of certain concepts and images may change throughout the assessment procedure.¹ Moreover, the context in which an assessor is asked to describe a media item influences the decision regarding which concepts he uses. Thus up to a point, relevance is subjective. Schamber (1994) identify 80 factors that affect relevance judgements, which Harter

¹Imagine the concept *cute*. After annotating a reasonable number of images, fewer images will occur cute to the assessor than in the beginning of the task.

(1996) sort into six groups: 1) characteristics of judges, 2) requests, 3) documents, 4) information systems, 5) judgement conditions, and 6) choice of scale. In text retrieval evaluation, this issue and its effects on the performance of systems have been systematically investigated, for example in Lesk and Salton (1968) and in Voorhees (2000). Results show that despite the large disagreement between assessors, the relative ranking of evaluated systems remains stable at least for assessors from the same population. Bailey et al. (2008) demonstrate that slight changes in the performance occurred averaged over a number of queries for assessors from different populations. Further, Lesk and Salton (1968) prove that the inter-assessor agreement is higher for top-ranked documents.

3.2.3.2. Completeness

In the Cranfield experiments, the assessors judged relevance for the entire collection. Cleverdon stressed the importance of having a complete set of judgements. Further, the computation of recall-based evaluation measures requires knowledge about the complete set of relevant documents. However, this is an intractable requirement with growing collection size. As a result, evaluation methods with incomplete judgements were investigated that still allow calculating recall-based performance measures. A popular method to reduce the number of necessary judgements is *pooling* as presented in Spärck Jones and van Rijsbergen (1975). Pooling is a special kind of sampling strategy. It uses the top k -ranked documents of each system for a topic and constructs a pool out of them by removing duplicates. The pooled documents are judged by relevance assessors without knowledge of which systems contributed to their being in the pool. Usually, all unjudged documents are regarded as irrelevant (Salton (1992)) or ignored in the performance calculation (Buckley and Voorhees (2004)). Pooling is based on the assumption that, for a sufficient number of systems and a sufficient depth of the pools, all relevant documents appear in the top k -ranked lists of at least one system. Usual numbers for k are 50 or 100 with a measurement depth of roughly 1000 documents. Some research was conducted concerning the question of whether pooling allows for a fair comparison of systems. With pooling, recall is overestimated as not all relevant documents are found, but the relative comparison between systems appears to be reliable. Zobel (1998) proposes a varying pool depth per topic that is determined by incrementing the depth size in small steps and estimating the likelihood of relevant documents to be found. With increasing size of test collections, Blair (2002) has raised the concern if pooling to a depth of 100 is still sufficient as the proportion of assessed to unassessed documents increases to 99%. Alternatives to pooling such as interactive search and judging of documents and focusing the assessors effort to particular systems are investigated by Cormack et al. (1998) and Aslam et al. (2003), respectively. Other research concentrates on evaluation without relevance judgements as to rank systems by reference count and with random relevance approaches (Efron (2009), Wu and Crestani (2003), and Soboroff et al. (2001)). Carterette (2008) focuses on how to limit the number of judgements needed with Minimal Test Collections. Two or more systems are compared with a minimum of judgements, while no effort is made to assess all relevant documents or to make a reliable estimate of evaluation measures.

3.2.3.3. Reusability

The reusability of judgements is closely related to the issue of completeness. Reasons to create reusable test collections lie mainly in the high costs of constructing and assessing collections. The costs are more likely to amortize if collections can be reused after the evaluation benchmark. Further, it is desirable to compare systems that were not evaluated in the benchmark with those from the benchmark. While judgements are reusable in cases where the whole collection was assessed, this picture can change with incomplete judgements. The problem lies in predicting what future systems might retrieve when building the pool for relevance assessment. A radical

new system may find relevant documents no other system has retrieved and that have therefore not been judged. Thus, runs that do not contribute to the pool are more likely to retrieve unjudged documents and are therefore potentially underestimated. Baillie et al. (2008) investigate the uncertainty that is brought into the test collection through incomplete judgements and propose a measure of uncertainty to decide whether a system comparison can be fairly made or not.

3.2.4. Databases

Multimedia databases are generally characterised by varying properties, be it the size of documents or the way they include metadata such as class labels or segment information. Although there are hundreds of different collections available, their use is often limited due to copyright restrictions that hinder the distribution of the collection. Further, a lot of collections are small and were build for a special purpose, commonly they cannot be made available by the authors. Databases should follow the requirements defined in Section 3.2.1 in order to be applicable in an evaluation scenario.

It is not within the scope of this work to provide an extensive list of available databases. However, the more important databases are listed. As this information is constantly changing, the reader is referred to a current effort of the CHORUS+ project that sets up a wiki of multimedia databases² or to the collection of the MUSCLE project³.

For image retrieval and annotation, the commercial Corel Database or subsets of it are often applied, although this database has been criticized in the community (see, for example in Müller et al. (2002)). Moreover, the evaluation initiative ImageCLEF has collected a great amount of databases for image retrieval and annotation that are summarised by Grubinger et al. (2010) (see also Section 3.4). Popular datasets for image retrieval and annotation, object detection, or segmentation are further the IAPR TC12 dataset (Grubinger et al. (2006)), the Caltech datasets (Griffin et al. (2007)), the PASCAL VOC (Everingham et al. (2010)), the Berkeley Segmentation Dataset (Martin et al. (2001)), the NUS-Wide database (Chua et al. (2009)), ImageNet (Deng et al. (2009) and Deng et al. (2010)), or the Scene UNderstanding (SUN) database (Xiao et al. (2010)). Huiskes and Lew (2008) have collected the MIR Flickr database with 25,000 freely distributable images from Flickr and later extended it to 1 million images (Huiskes et al. (2010)). In video retrieval, the TRECVID databases originated by the TRECVID benchmark (Smeaton et al. (2006)) are established as a gold standard and are widely accepted in the research community. Large-scale collaboratively annotated collections are available from Russell et al. (2008) for image sets and Yuen et al. (2009) for video sets.

3.3. Evaluation measures

Evaluation measures are the means of determining the performance of a system on a defined test collection. During the years of IR and VIR research, many different evaluation measures have been proposed, as summarised in the works from van Rijsbergen (1979), Manning et al. (2009), and Demartini and Mizzaro (2006). Their applicability is determined by the evaluation objective and the design of the test collection. This section focuses on evaluation measures adopted for the evaluation of multi-label classification of media items. The shift from single-label categorisation to multi-label classification in image annotation allows applying two evaluation paradigms. The first directly evolves from categorisation evaluation, in which the evaluation per category is sufficient as each media item exclusively belongs to one class. Performance is assessed for each concept in isolation and averaged by the total number of concepts. The second paradigm starts with the media item and evaluates whether all concepts are assigned correctly on a per item basis. Then,

²<http://www.ist-chorus.org/wiki/index.php/Category:Dataset>, last visited 25.06.2011

³https://muscle.caa.tuwien.ac.at/data_links.php, last visited 25.06.2011

instead of comparing a single predicted label to a single ground truth label, one needs to compare two sets of labels. As a result, the predicted labels can be *fully correct* (label sets are identical), *fully wrong* (the intersection of the sets is empty), or *partly correct* (the sets have common labels, but are not fully identical).

Multi-label classification evaluation can follow these two evaluation paradigms. Tsoumakas and Vlahavas (2007) name them *example-based evaluation* and *label-based evaluation*. The example-based evaluation generates a score for each media item (example) and then averages by all items. The label-based evaluation subdivides the annotations in a single evaluation per concept and averages by all concepts. Throughout this work, the latter is called *concept-based evaluation* to distinguish it from the word *label* that is used for the concrete annotations for a media item.

Further, the way in which relevance is assessed by the assessors and by the annotation systems influences the choice of the evaluation measure. Judges can assess binary or graded relevance to a media item as discussed in Section 3.2.3. Graded relevance requires an evaluation measure that incorporates the different grades into the evaluation function. A popular evaluation measure for graded relevance is the Discounted Cumulative Gain proposed by Järvelin and Kekäläinen (2002). Evaluation measures that are able to cope with graded relevance are not discussed further here, as multi-label classification scenarios usually consider binary relevance (see Chapter 4.3). Second, the annotations retrieved by the VIR systems can be binary decisions or likelihoods of how certain a system is about the annotation of a particular concept. Evaluation measures for label set predictions can be applied to confidence outputs of a VIR system after a threshold maps the likelihoods to a binary decision. However, often systems that use label set predictions and are evaluated with evaluation measures for ranked annotations are disadvantaged. Demartini and Mizzaro (2006) perform a classification of different effectiveness measures for retrieval, based on the year of appearance in literature and the kind of relevance and retrieval types the measures rely on (binary relevance, ranked relevance, or continuous relevance, respectively retrieval). In total, 44 measures are classified that cover all measures published between 1965 and 2005 that were applied to IR research.

In the following, the author introduces the most important evaluation measures for label set and ranked predictions considering the concept-based and example-based evaluation.

3.3.1. General notations

In multi-label classification, each data example (in this case each image) is associated with a set of labels. Let X be a dataset consisting of examples X_n , $n = \overline{1, N}$, where N is the total number of examples in the dataset. The class membership for the dataset X is denoted as $Y = \{y_{nc}\}$, $c = \overline{1, C}$, where C is the total number of concepts, and

$$y_{nc} = \begin{cases} 1, & \text{if example } n \text{ belongs to concept } c, \\ 0, & \text{otherwise.} \end{cases} \quad (3.1)$$

In other words, the ground truth annotation matrix $\{y_{nc}\}$ is a binary matrix, where the rows y_n correspond to examples and the columns y_c correspond to concepts. In each matrix row y_n , each non-zero element y_{nc} indicates that the example n is associated with a concept c . In the case of multi-label classification, each row of Y can have multiple non-zero values. Likewise, the non-zero elements of the column y_c indicate the examples belonging to the concept c . If it is not stated otherwise, the ground truth annotations are denoted as Y , \mathcal{Y}_n , y_c , or y_{nc} , while the estimated annotations (suggested by the system) are denoted as Z , \mathcal{Z}_n , z_c , or z_{nc} , respectively. The predicted annotations are binary values or floating point values denoting confidences in the range of $[0 : 1]$. With $f(z_{nc})$, the confidence score assigned to concept c for item n is denoted.

Characteristics of multi-label datasets: The important characteristics of a multi-labelled dataset are label cardinality (LC) and label density (LD). LC shows how many labels have been assigned to a media item in average:

$$LC(X) = \frac{1}{N} \sum_{n=1}^N \sum_{c=1}^C y_{nc}, \quad (3.2)$$

and the LD is defined as fraction of the average number of the used labels to the total number of available labels:

$$LD(X) = \frac{1}{N \cdot C} \sum_{n=1}^N \sum_{c=1}^C y_{nc}. \quad (3.3)$$

3.3.2. Concept-based evaluation measures for label set predictions

Concept-based evaluation measures judge the quality of annotation systems for each concept. In the following, different concept-based evaluation measures are introduced. These measures need a binary decision of the VIR system about the presence and absence of concepts.

Precision, Recall, F-Measure, Accuracy: The traditional information retrieval evaluation measures, namely Precision, Recall, F-measure, and Accuracy are initially calculated for each concept independently. For the concept c , two binary vectors z_c and y_c are compared. The number of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) examples is calculated for the concept c . If the elements of the binary vectors are treated as logical values, then $TP(c)$, $FP(c)$, $FN(c)$ and $TN(c)$ can be written as:

$$TP(c) = \sum_{n=1}^N (z_{nc} \wedge y_{nc}), \quad (3.4)$$

$$FP(c) = \sum_{n=1}^N (z_{nc} \wedge \neg y_{nc}), \quad (3.5)$$

$$FN(c) = \sum_{n=1}^N (\neg z_{nc} \wedge y_{nc}), \quad (3.6)$$

$$TN(c) = \sum_{n=1}^N (\neg z_{nc} \wedge \neg y_{nc}). \quad (3.7)$$

Then, the concept-based Precision, Recall, F-measure, and Accuracy are defined as follows:

$$\text{Precision}_{cb}(Z, Y) = \frac{1}{C} \sum_{c=1}^C \frac{TP(c)}{TP(c) + FP(c)}, \quad (3.8)$$

$$\text{Recall}_{cb}(Z, Y) = \frac{1}{C} \sum_{c=1}^C \frac{TP(c)}{TP(c) + FN(c)}, \quad (3.9)$$

$$\text{F-measure}_{cb}(Z, Y) = \frac{1}{C} \sum_{c=1}^C \frac{2 \cdot TP(c)}{2 \cdot TP(c) + FP(c) + FN(c)}, \quad (3.10)$$

$$\text{Accuracy}_{cb}(Z, Y) = \frac{1}{C} \sum_{c=1}^C \frac{TP(c) + TN(c)}{TP(c) + FP(c) + FN(c) + TN(c)}. \quad (3.11)$$

These equations use *macro-averaging* over all concepts. In contrast, *micro-averaging* would consider the averages of TP, FP, FN, TN separately for all concepts before the evaluation function is applied. Micro-averaging gives equal weight to every document in the collection, while macro-averaging gives equal weight to every concept (Tague-Sutcliffe (1992)). See Equation 3.12 as an example for the micro-averaged concept-based precision:

$$\text{Precision}_{cb_{micro}}(Z, Y) = \frac{\sum_{c=1}^C TP(c)}{\sum_{c=1}^C TP(c) + \sum_{c=1}^C FP(c)}. \quad (3.12)$$

3.3.3. Concept-based evaluation measures for ranked assignments

Precision, Recall, F-Measure, and Accuracy are calculated based on binary estimates of the system concerning whether a concept should be assigned or not as illustrated in the preceding section. In this section, some concept-based evaluation measures that rely on likelihoods of the system are introduced. The likelihoods are used to rank results for a particular query and the measures take the positions in the ranked lists into account to calculate the score.

AUC, EER - ROC curve measures: The concept-based measures AUC and EER can be calculated from the Receiver Operating Characteristics (ROC) curve and are common measures for different recognition tasks, for example in Liu and Shriberg (2007). A ROC curve is a graphical plot of the true-positive-rates against the false-positive-rates of a binary classifier according to different threshold values. The EER is defined as the point where the errors from assigning images to a concept equal the errors from not assigning them to the concept. The measure AUC describes the overall quality of a classification system independently from an individual threshold configuration, with the specific trade-off between TPs and FPs. It is calculated by integration of the ROC curve, whereas an AUC value of 1 equals a perfect system with no FPs, and an AUC value of 0.5 equals a random system.

Average Precision: Average Precision (AP) is a concept-based measure that approximates the area below the uninterpolated Precision-Recall Curve (PR-curve). In other words, it is the average of the Precision values calculated after each relevant document is retrieved for a single query. Usually, it is averaged over several information needs and then called MAP. MAP is calculated as

$$MAP = \frac{1}{C} \cdot \sum_{c=1}^C \frac{\sum_{rn=1}^N (P(rn) \times rel(rn))}{R_n}, \quad (3.13)$$

with rn as rank number; $rel(rn)$ denoting the relevance (1 or 0) of the item; $P(rn)$ is the precision at item rn and R_n is the total number of relevant documents for a concept. MAP is often used as a single-value measure that summarises the quality across recall levels of ranked retrieval results and is, for example, utilised as standard evaluation measure in the TREC community. For a detailed explanation, see Manning et al. (2009).

A similar measure to AP is the interpolated Average Precision (iAP). Instead of calculating the Precision after each relevant document is retrieved, it calculates Precision on a number of defined Recall levels. Usually, eleven Recall levels (0%, 10%, ..., 100%) are chosen. Averaged over several information needs, this is called Mean interpolated Average Precision (MiAP).

Yilmaz and Aslam (2006) propose how to infer AP when dealing with incomplete and imperfect judgements. The inferred Average Precision (infAP) estimates AP and simulates the results as if they would have been evaluated with complete relevance judgements. In further work, Yilmaz et al. (2008) propose the extended inferred Average Precision (xinfAP) which incorporates non-random relevance judgements.

Other concept-based evaluation measures for ranked predictions: Precision at a fixed rank and R-precision are other popular evaluation measures. While the first calculates the precision at a predefined rank, the latter calculates precision at the rank that equals the number of relevant documents for a certain concept. R-Precision is sometimes also referred to as break even point precision (BEP). The reader is referred to Sanderson (2010) and Manning et al. (2009) for further information.

3.3.4. Example-based evaluation measures for label set predictions

Example-based evaluation measures judge the quality of annotation systems for each media item. They consider the full set of labels for a media item and average over the number of items. In the following, different example-based evaluation measures that rely on binary system predictions are introduced.

Precision, Recall, F-Measure, Accuracy: In contrast to the concept-based variants of Precision, Recall, and F-measure, the scores are first calculated for each example and then averaged over all examples. Precision, Recall, F-measure, and Accuracy are defined as follows:

$$\text{Precision}_{eb}(Z, Y) = \frac{1}{N} \sum_{n=1}^N \frac{|\mathcal{Y}_n \cap \mathcal{Z}_n|}{|\mathcal{Z}_n|}, \quad (3.14)$$

$$\text{Recall}_{eb}(Z, Y) = \frac{1}{N} \sum_{n=1}^N \frac{|\mathcal{Y}_n \cap \mathcal{Z}_n|}{|\mathcal{Y}_n|}, \quad (3.15)$$

$$\text{F-measure}_{eb}(Z, Y) = \frac{1}{N} \sum_{n=1}^N \frac{2 * |\mathcal{Y}_n \cap \mathcal{Z}_n|}{|\mathcal{Y}_n| + |\mathcal{Z}_n|}, \quad (3.16)$$

$$\text{Accuracy}_{eb}(Z, Y) = \frac{1}{N} \sum_{n=1}^N \frac{|\mathcal{Y}_n \cap \mathcal{Z}_n|}{|\mathcal{Y}_n \cup \mathcal{Z}_n|}. \quad (3.17)$$

Hamming Loss: Hamming Loss is the average proportion of misclassified labels. It counts the FP and FN predictions of the system. The smaller the value for Hamming Loss, the better the performance.

$$\text{Hamming Loss}(Z, Y) = \frac{1}{N} \sum_{n=1}^N \frac{1}{C} |\mathcal{Z}_n \Delta \mathcal{Y}_n|. \quad (3.18)$$

Here, Δ denotes the symmetric set difference. In Boolean logic terminology, Δ equals the *XOR* operation.

Alpha evaluation: Shen et al. (2004) propose an α -evaluation and multi-label class Recall and Precision. α -evaluation generates a score while taking the ground truth, predicted labels, FNs, and FPs into account. Moreover, FP and FN labels can be penalised differently if it is more suitable for the particular application. The parameter α introduces the so-called *forgiveness rate* as a trade-off between the fully correct and partly correct prediction. Equation 3.19 depicts the α -evaluation formula for equally treated FPs and FNs:

$$\text{score}(\mathcal{Z}_n, \mathcal{Y}_n) = \left(\frac{|\mathcal{Y}_n \cap \mathcal{Z}_n|}{|\mathcal{Y}_n \cup \mathcal{Z}_n|} \right)^\alpha \quad \alpha \geq 0 \quad (3.19)$$

$$\text{Accuracy}_\alpha(Z, Y) = \frac{1}{N} \sum_{n=1}^N \text{score}(Z_n, Y_n) \quad (3.20)$$

If $\alpha = 1$, Equation 3.20 is equal to the example-based Accuracy measure.

3.3.5. Example-based evaluation measures for ranked assignments

A few example-based evaluation measures that consider confidence values for calculating the score are summarised in Schapire and Singer (2000) as well as in Zhang and Zhou (2007). So far, they are not established as evaluation measures in IR and VIR evaluation. All measures presented consider confidences to rank the labels assigned to a media item.

One-Error: One-Error computes the proportion of images for which the top-ranked label is not in the true set of concepts. We define $\beta_n = \max(z_{n1}, z_{n2}, \dots, z_{nc})$. The class c which satisfies this formula for a given image n is further denoted as c_n . Then One-Error is computed as follows:

$$\begin{aligned} \text{One-Error}(Z, Y) &= \frac{1}{N} \sum_{n=1}^N f(c_n), \\ \text{with } f(c_n) &= \begin{cases} 1, & c \notin y_n, \\ 0, & c \in y_n. \end{cases} \end{aligned} \quad (3.21)$$

Its value is between 0 (best performance) and 1 (worst). The measure One-Error ignores all assigned labels at rank 2 or below.

Coverage: Coverage computes how far, on average, one has to go down in the ranked list of labels beyond the number of true concepts to cover all concepts from the image's ground truth.

$$\text{Coverage}(Z, Y) = \frac{1}{N} \sum_{n=1}^N \max_{c \in Y_n} \text{rank } f(z_{nc}) - |y_n|, \quad (3.22)$$

with rank as the ranking function that sorts the confidence values in Z in descending order. Smaller values describe better systems: $\text{Coverage}(Z, Y) = 0$ is equivalent to perfect algorithmic ranking of concepts. I have modified the formula given in the literature (Schapire and Singer (2000)) by subtracting $|y_n|$ instead of 1 to make the score for perfect performance independent of the number of true concepts. Coverage depends on the worst-ranked true concept for each image. If the number of concepts is large, then a few incorrect low label ranks can unduly increase the value for coverage.

Ranking Loss: The measure Ranking Loss calculates the average fraction of how often an irrelevant concept is ordered before a relevant one in the ranked prediction:

$$\begin{aligned} \text{Ranking Loss}(Z, Y) &= \frac{1}{N} \sum_{n=1}^N \frac{|f(z_{nc}) \leq f(z_{nc'})|}{|R(Y_n)| |R(\overline{Y_n})|}, \\ &\text{with } (c, c') \in R(Y_n) \times R(\overline{Y_n}). \end{aligned} \quad (3.23)$$

$R(\overline{Y_n})$ corresponds to the set of non-relevant concepts in Y_n and is defined as $Y_n \setminus R(Y_n)$. The theoretically best performance has a ranking loss value of 0.

Average Precision: The non-interpolated AP can also be defined for the evaluation per example. In this case, not the ranking of documents, but the ranking of concepts per document is considered to calculate the score.

$$MAP_{ex} = \frac{1}{N} \cdot \sum_{n=1}^N \frac{\sum_{rc=1}^C (P(rc) \times rel(rc))}{R_c}, \quad (3.24)$$

with rc as rank number; $rel(rc)$ denoting the relevance (1 or 0) of the concept; $P(rc)$ is the precision at concept rc and R_c is the total number of relevant concepts for a document.

R-Precision: The measure R-Precision measures the Precision at the rank of perfect Recall. It can also be defined for the example-based evaluation and is then computed as follows:

$$\text{R-Precision}_{ex}(Z, Y) = \frac{1}{N} \sum_{i=1}^N \frac{|\mathcal{Y}_n \cap \mathcal{Z}_n^*|}{|\mathcal{Y}_n|}, \quad (3.25)$$

with $\mathcal{Z}_n^* \subseteq \mathcal{Z}_n \Leftrightarrow \forall c \in \mathcal{Z}_n : \text{rank}(f(z_{nc})) \leq |\mathcal{Y}_n|$.

3.3.6. Significance tests

In VIR evaluation, the performance of a system is often compared to the performance of other systems. The relative differences between all systems evaluated on the same test collection determine the benefits of each approach. Significance tests allow investigating whether the difference between two systems is due to a better approach, or if it may result from random chance. They are thus a means of estimating the reliability of the evaluation results.

Hypothesis testing estimates the probability p of observing a certain difference in performance, given that the *null hypothesis* H_0 is true. In VIR evaluation, H_0 states that two systems are in effect the same and that any difference is due to random chance. Following, if p is below a certain threshold (normally $p < 0.05$ or $p < 0.01$), it is concluded that H_0 is unlikely and should be rejected. The alternative hypothesis, namely that the systems show significant differences, is then assumed to be true. Significance tests can produce errors that are categorised into Type I and Type II errors. Type I errors occur when a significance test incorrectly rejects H_0 , while Type II errors occur if a significance test incorrectly concludes that H_0 cannot be rejected. Depending on the *power* of the significance test, more Type I or Type II errors are likely to result. Tests with fewer assumptions, such as non-parametric significance tests, tend to produce more Type II errors, whereas parametric tests with more underlying assumptions tend to produce Type I errors. See Sanderson (2010), Chapter 5 for more information on different types of significance tests.

3.4. Benchmarks for multimedia retrieval

Due to the variety of database characteristics, results of retrieval algorithms are often incomparable and it is hardly possible to decide whether one approach outperforms another or not. As a result, several benchmarking activities for different tasks in multimedia retrieval have become popular. Benchmarks define challenging tasks with the goal of objectively measuring the performance of algorithms and of establishing a baseline for comparing systems. When defining the tasks of a benchmark, criteria such as objectivity, scalability, processing times, user's interest, and expected real-world scenarios are important to consider. The main goal of benchmarks is to advance the research field by the means of a standardised test collection and to build up a community on the research topic. Annual workshops help to discuss and present results, and inspire research directions. In the long run, this generates methods that work.

The history of IR benchmarking begins with TREC. In 1990, the National Institute of Standards and Technology (NIST) was funded to build a large test collection for text retrieval research, which

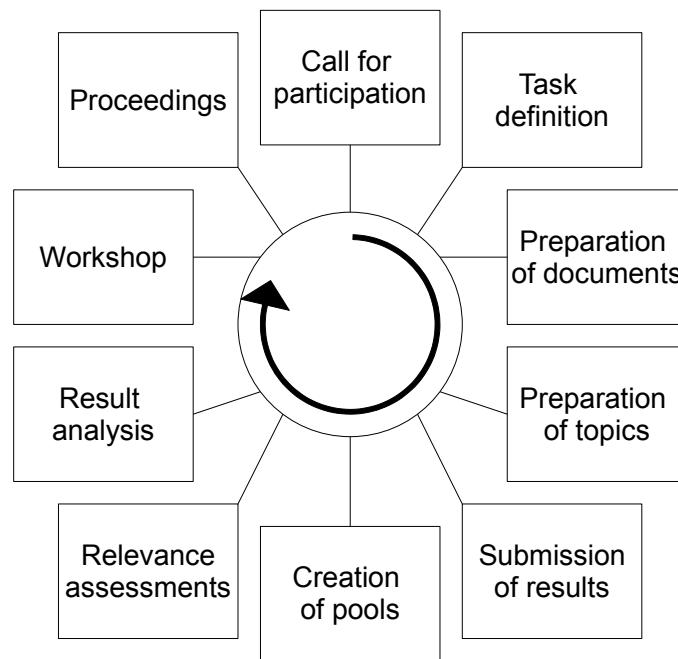


Figure 3.1.: Evaluation cycle for TREC-style evaluation benchmarks (adapted from <http://trec.nist.gov/presentations/TREC2004/04intro.pdf>)

was made available to the community in 1991. TREC started with its first annual evaluation event in 1992 (Voorhees and Harman (2005)) and set the stage for all following IR benchmark initiatives. It took 9 years from the start of TREC until evaluation benchmarks arrived in Multimedia Information Retrieval (MIR) research. TRECvid started in 2001 as one track of TREC and evolved to an independent benchmarking initiative in 2003 (Smeaton et al. (2006)). It focuses on video retrieval evaluation, as shot-boundary detection, ad-hoc search, high-level feature extraction, or rushes summarisation. ImageCLEF was initiated in 2003 as a track of Cross Language Evaluation Forum (CLEF) and is devoted to the evaluation of cross-language retrieval in image collections. In its first evaluation cycle, it hosted two tasks, photographic retrieval and interactive image retrieval, and evolved to an active track with varying tasks and a registration of 112 international research groups in 2010. In 2005, the PASCAL VOC challenge started as evaluation benchmark for object recognition in images (Everingham et al. (2010)). The Music Information Retrieval Evaluation eXchange (MIREX) benchmark is concerned with music retrieval evaluation (Downie (2008)) and started in 2005. The Russian Information Retrieval Evaluation Seminar (ROMIP)⁴ evaluation seminar focuses on Russian language text retrieval since 2003 and has hosted a multimedia track on image retrieval since 2008. VideoCLEF ran as video evaluation track in CLEF 2008 and 2009, and launched as MediaEval⁵ a separate evaluation initiative in 2010. An overview on multimedia benchmarks is summarised in Little et al. (2010).

Most IR benchmarks follow the Cranfield paradigm in their evaluation process, although there are a few interactive retrieval tasks that focus on user-centred evaluation. Usually, benchmarks are defined through a cycle of activities, often lasting for one year. The stations in the benchmark cycle are illustrated in Figure 3.1. The benchmark starts with a call for participation in the beginning of the yearly cycle. Research groups express their interest in participating and register for the tasks they plan to solve. Sometimes, this step also involves signing copyright agreements in order to

⁴<http://romip.ru/en>

⁵<http://www.multimediaeval.org>

get access to the test collections. Next, the tasks are defined in detail, naming the evaluation objective, the test collection, and the evaluation measures that will be applied. Depending on the task, training data may also be released in this stage. In the following stages, the documents of the test collection and the topics are released. The participants apply their systems, which may be tuned on the training set beforehand, to the topics and generate one or more result files for the task. These result files are called *runs* and are submitted to the task organisers. The task organisers use the runs to create pools and perform the relevance assessments. Note that fully assessed collections can be finalised prior to the submission of runs. The submissions of the participants are evaluated and analysed. In the next stage, the participants and task organisers meet in a workshop to discuss the results and tasks, and to decide on challenging objectives for the next evaluation iteration. Finally, the results are published in the workshop proceedings.

Organisers of benchmarks have to face some challenges. According to Müller et al. (2007), the biggest organisational challenges lie in the funding, in the access to datasets which are often copyrighted, in the motivation of participants, in the relevance assessments, and in the creation of realistic tasks and user models. Despite the highly visible benefits benchmarking has brought to the research community, there also exists criticism. The main critics address the fear of a standardisation of research, the tendency to reuse techniques known to perform well, and the question of whether the evaluation without involvement of end users makes sense at all. Also, benchmarks are sometimes seen as competitions rather than as platforms to test new ideas and to become involved in discussions with the community. Benchmark organisers try to counteract these tendencies by changing task objectives and test collections, including feedback, and motivating people to organise new tasks.

3.5. Summary

This chapter provided an in-depth review of evaluation in IR with a special focus on VIR evaluation. It began with a brief overview of the development of experimental evaluation. Subsequently, the parts of test collections were introduced and discussed. In particular, the difficulty to define concepts based on a user model and to determine relevance was highlighted. The most important performance measures for multimedia annotation evaluation were introduced. Finally, multimedia benchmarks were presented including a description of their usual evaluation cycle. This overview on evaluation in IR sets the stage for the following chapters. A detailed review of related work for special subtopics on image annotation evaluation is presented in the next chapter. It directly focuses on works related to the proposals in this thesis.

4. Related work on evaluation methodologies for image annotation



The previous two chapters have introduced the process of multi-label image annotation and evaluation. The main contributions of this thesis are the definition of test collections for image annotation and evaluation methodologies for image annotation evaluation. The general idea is based on the lack of evaluation methodologies for performance measurement of image annotation, as current evaluation initiatives just adopt principles from IR evaluation without considering that multi-label image annotation poses its own challenges. This chapter reviews related work dealing with aspects relevant for the definition of test collections and performance measures considering semantics. The chapter is structured as follows. Section 4.1 summarises the use of visual concepts in test collections. Then, Section 4.2 reviews related work on crowdsourcing for relevance assessment in IR. Following this, Section 4.3 defines dimensions of IR performance measures and analyses the usage of performance measures in image annotation literature. Section 4.4 details related work on performance measures considering semantics, while Section 4.5 introduces semantic relatedness measures. Finally, Section 4.6 presents an outlook on how to evaluate new performance measures. Each section ends with a discussion on unresolved points in related work and on how these points will be tackled in this thesis.

4.1. Visual concepts for image annotation

Visual concepts denote one part of an image test collection. A brief introduction on the history of visual concepts in VIR test collections was given in Chapter 3.2.2, while a summary of visual concepts in early datasets can be found in Hanbury (2007). In contrast, this section focuses on concepts of recent datasets that are commonly used in image annotation. Many recent datasets provide a structuring of concepts in hierarchies or ontologies. These are presented if available.

The test collections of the [PASCAL VOC](#) classification challenge (Everingham et al. (2010)) are often applied in image annotation evaluation. They define 20 object classes which are not further structured. The [MIR Flickr test collection](#) proposed by Huiskes and Lew (2008) defines 23 visual concepts derived from the Flickr user tags of the collection. The concepts are structured

into general topics and subtopics. The general topics include the concepts *sky, water, people, night, plant life, animals, man-built structures, sunset, indoor, and transport*. The 81 visual concepts of the NUS-WIDE collection (Chua et al. (2009)) are derived from frequent Flickr tags and related work. The authors use the categories *scene, object, event, program, people, and graphics* as main categories to structure the concepts. Lin et al. (2003) have categorised 133 concepts of the VideoAnnEx test collection in the categories *event, scene, and objects*. These 133 concepts are designed for a video corpus and consist of audio-visual events, visual scenes, sounds, and visual objects. The Caltech-256 dataset (Griffin et al. (2007)) includes 256 object categories and uses a manually created taxonomy which differentiates between *animate* and *inanimate* objects at the top level. Escalante et al. (2010) propose the segmented and annotated IAPR TC-12 test collection. They define 274 concepts that are derived from concepts in related work and adapted to the images of the collection. A manually created hierarchy divides the concepts into the six top level categories *animal, landscape, man-made, human, food, and other*. Xiao et al. (2010) propose the SUN database of 899 scene concepts, including a subset of 397 well-sampled concepts. The scenes are selected from the 70,000 terms of WordNet by choosing all terms that describe scenes, places, and environments. The concepts are structured into two category levels. The first level divides concepts into the categories *indoor, outdoor-natural, and outdoor-man-made*. These are further subdivided on the second level. The construction of the LSCOM dataset and ontology (Naphade et al. (2006)) considers user input, critiques from experts, comparison to related ontologies, and evaluation based on utility, observability, and feasibility criteria. The 856 concepts included in LSCOM version 1 are structured concerning the categories *activities/events, objects, scenes/locations, people, graphics, and programs* and are focused on video annotation. The ImageNet database uses the noun synsets of WordNet as categories.¹ In total, WordNet consists of around 80,000 noun synsets which will be filled with images in the long-term goal of ImageNet. In Deng et al. (2009), ImageNet is described as consisting of 12 subtrees containing 5,247 synsets. According to the website², this has expanded to 27 subtrees containing 17,624 synsets. The current top level categories are *amphibian, animal, appliance, bird, covering, device, fabric, fish, flower, food, fruit, fungus, furniture, geological formation, invertebrate, mammal, musical instrument, plant, reptile, sport, structure, tool, tree, utensil, vegetable, vehicle, and person*.

4.1.1. Discussion

The number and structure of visual concepts are diverse in the different test collections. Current datasets range from small sets of about 20 concepts to large-scale Web datasets with about 17,500 concepts. Some datasets provide no structure at all (PASCAL VOC), while others come with a complete ontology (LSCOM) or are even built upon an existing ontology (ImageNet). Further, the datasets often show no real multi-label characteristics, as the concepts are usually first defined and then images are sought for each concept category. As a result, the visual concept is often depicted in the middle of the image and often, no other concept is visible. This is especially true for the object-based datasets as Caltech-256. Moreover, this unifies the number of images per concept, which differs from the real-world case.

The definition of visual concepts and a new ontology for the annotation of consumer images will be presented in Chapter 5. A special focus is placed on an all-embracing view on how to index images, to unify different views such as objects, scenes, or representational issues, and on including user needs. Further, the images of the MIR Flickr collection are used as document base. These images were collected before the definition of visual concepts has started, which ensures the generality of the image content.

¹The word "synset" stands for "synonym set" and describes English terms that are grouped together as synonyms.

²<http://www.image-net.org/about-stats>, last visited 31.05.2011

4.2. Relevance assessment adopting crowdsourcing

Test collections allow comparing the effectiveness and quality of systems in system-centred evaluation. The creation of large, semantically annotated corpora from scratch is a time- and cost-consuming activity, especially because of the required relevance judgements. Section 3.2.3 introduced the process of relevance assessment in IR with its crucial issues of consistency, completeness, and reusability. Normally, the relevance assessment is performed by experts from the field of interest (for example, in evaluation of medical image annotation, usually physicians assess the relevance of images for a certain medical topic). But there are fields in which no expert knowledge is needed to decide on relevance, and in which the assessment can be performed with common-sense knowledge. Recently, different works have been presented that outsource multimedia annotation tasks to crowdsourcing approaches. According to Howe (2006),

crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call.

Often, the work is divided into small parts and distributed to a large community over Web-based platforms. Utilising crowdsourcing approaches for assessing ground truth corpora is mainly motivated by the reduction of costs and time. Other motivational factors include diversity in opinion and diversity in population.

4.2.1. Crowdsourcing at Amazon Mechanical Turk

Most crowdsourcing approaches adopt the online marketplace Amazon Mechanical Turk (MTurk)³ which allows distributing mini-jobs to a crowd of people. At MTurk, these mini-jobs are called Human Intelligence Tasks (HITs). They represent a small piece of work with an allocated price and completion time. The workers at MTurk, called turkers, can choose the HITs they would like to perform and submit the results to MTurk. The requester of the work collects all results from MTurk after they are completed. The workflow of a requester can be described as follows: 1) design a HIT template, 2) distribute the work and fetch results, and 3) approve or reject work from turkers. For the design of the HITs, MTurk offers support by providing a Web interface, command line tools, and developer Application Programming Interfaces (APIs). The requester can define how many assignments per HIT are needed, how much time is allotted to each HIT, and how much to pay per HIT. MTurk offers several ways of assuring quality. Optionally, the turkers can be asked to pass a qualification test before working on HITs, multiple workers can be assigned the same HIT, and requesters can reject work in case the HITs were not finished correctly. The HIT approval rate each turker achieves by completing HITs can be used as a threshold for authorisation to work. Moreover, the completion times provide information on the reliability of the answers of turkers. For more information on MTurk, also see Ipeirotis (2010) and Chen et al. (2011). The interested reader is referred to Ross et al. (2010) to learn about demographics of turkers.

Crowdsourcing is a rather new technology for which no standardisation of approaches yet exists. The main challenges lie in the HIT design, in how to balance payments (Feng et al. (2009), Mason and Watts (2009), Horton and Zeckhauser (2010)), in ethics regarding the possible exploitation of workers in third party countries (Fort et al. (2011)), in legal issues (Felstiner (2010)), and in quality control of the obtained answers. As this thesis focuses on crowdsourcing for relevance judgements, the following discussion is limited to quality control and consistency of relevance judgements and to crowdsourcing for visual test collections.

³www.mturk.com

4.2.2. Quality assurance of relevance judgements

As explained in Section 3.2.3, consistency of judgements is a general issue in VIR and annotation evaluation. Several experiments have been performed on the analysis of inter-annotator agreements in order to assess the subjectivity in the acquisition process of ground truth for IR evaluation. Voorhees (2000) analyses the influence of changes in relevance judgements on the evaluation of retrieval results utilising the Kendall τ correlation coefficient (see Section 4.6.1). Volkmer et al. (2007) present an approach that integrates multiple judgements in the classification system and compare them to the kappa statistics. Brants (2000) proposes a study about inter-annotator agreement for part-of-speech and structural information annotation in a corpus of German newspapers. He uses the accuracy and F-score between the annotated corpus of two annotators to assess their agreement. A few studies have been performed to examine the inter-annotator agreement for word sense disambiguation (Véronis (1998), Chklovski and Mihalcea (2003)). These studies often utilise kappa statistics for calculating agreement between judges.

Despite the general subjectivity in performing relevance judgements, quality assurance is even more difficult to control in crowdsourced relevance judgements. These judgements can also be performed randomly by spammers or without a real understanding of the task. Approaches have to be investigated that are capable of detecting malicious workers and cheating. Some studies that explore the annotation qualities obtained with crowdsourcing approaches have been conducted. One of the first works on crowdsourcing for relevance assessment in IR was performed by Alonso et al. (2008). They propose crowdsourcing as alternative to editorial relevance assessment and introduce an experiment on relevance assessment of query-result pairs from turkers, including qualification tests. Later work by Alonso and Mizzaro (2009) examines how well relevance judgements for the TREC topic about space program can be fulfilled by workers at MTurk. The relevance of a document had to be judged regarding this topic, and the authors compared the results of the non-experts to the relevance assessment of TREC. They learned that the annotations among non-expert and TREC assessors are of comparable quality. Snow et al. (2008) investigate the annotation quality for non-expert annotators in five natural language tasks. They found that a small number of non-expert annotations per item yields to performance equal to that of an expert annotator. Moreover, they propose to model the bias and reliability of individual workers in order to derive an automatic noise correction algorithm. Hsueh et al. (2009) compare the annotation quality of sentiment in political blog snippets using a crowdsourcing approach and expert annotators. They define three criteria, the noise level, the sentiment ambiguity, and the lexical uncertainty, that can be used to identify high quality annotations. Kazai and Milic-Frayling (2009) examine measures to obtain the quality of collected relevance assessments. They point to several issues, such as topic and content familiarity, dwell time, agreement, or comments of workers that can be used to derive a trust weight for judgements. Grady and Lease (2010) investigate the influence of HIT design on the accuracy, cost, and time to assess document relevance for search tasks. They introduce variations in the query, the terminology, the payment, and the pay of bonus rewards. Eickhoff and de Vries (2011) analyse how tasks should be designed to distract spammers. They prove that malicious workers are less frequently working on novel tasks that require a certain degree of creativity, and recommend to design tasks in a non-repetitive way. Other work concerns itself with how to verify crowdsourced annotations (Chen et al. (2009)), how to deal with random annotations (Sheng et al. (2008), Donmez et al. (2009)), and how to estimate the quality of a worker (Ipeirotis et al. (2010)).

4.2.3. Crowdsourcing for image annotation

Most works presented deal with relevance assessment for textual collections. Related work on crowdsourcing for image annotation and on the reliability of these judgements is rarely investigated

for paid incentives. Several works exist that use human computation power (Yuen et al. (2009)) for labelling images such as LabelMe (Russell et al. (2008)), or games with a purpose such as the ESP Game (von Ahn and Dabbish (2004)). These offer other incentives, e.g., the entertainment component of games to reward voluntary effort. The following discussion is limited to work on paid crowdsourcing for image annotation. Please note that, to my knowledge, there exists no study that presents an extensive evaluation of the quality of these annotations in comparison to expert annotations and of the implications inherent in using a crowdsourced collection in benchmarks.

Sorokin and Forsyth (2008) were first in using crowdsourcing for image annotation with paid rewards. The workers were asked to perform an image segmentation and labelling task. Quality was assured by inclusion of gold answers, multiple judgements, and validation through other turkers (so-called grading), and used to filter occasional errors and to detect cheating. Deng et al. (2009) present the database ImageNet, which results from an ambitious initiative that aims at collecting a total of 50 million images annotated using the WordNet hierarchy. They collect candidate images from the Internet for each synset of WordNet by querying several image search engines in several languages. Each candidate image is verified by turkers who were provided with the definition of the synset. Multiple judgements were used to assure quality. Vijayanarasimhan and Grauman (2009) propose an active learning algorithm for estimating effort and information gain of a candidate image annotation. Endres et al. (2010) present challenges on collecting a detailed object recognition database including polygons, segmentation masks, and segment labelling using MTurk. Most issues result from the quality of the annotators (motivation, language skills, attention) and the complexity (ambiguity, confusion, needed accuracy) of the task. Soleymani and Larson (2010) assess boredom in videos with crowdsourcing. They follow a two-step approach in which turkers had to pass a pilot task with acceptable results before they were invited to participate in the main task. Yan et al. (2010) present a system for mobile image search, in which an automated search component ranks candidate images for a query. A predictive algorithm decides which results need to be validated. The validation is then performed by a total of five turkers per search result by relying on a majority vote for the final decision. Welinder and Perona (2010) present an approach that dynamically decides how many workers are needed per task in image labelling to obtain a reliable ground truth from multiple eventually contradicting labels. The algorithm provides an online estimate of the reliability of a worker in order to optimise resources for different types of annotations (binary, multi-valued, continuous).

4.2.4. Discussion

All of these studies employ methods for quality estimation and base a rejection of work on these results. However, a reliable rejection mechanism is difficult to set up, as the task of image annotation entails ambiguities and subjectivity. The construction of test collections with crowdsourcing aims at the cheap acquisition of relevance judgements in large-scale collections which can be used to evaluate annotation systems. An analysis of the noise that is brought into the evaluation process due to inconsistencies in relevance judgements and its effect on system ranking is still lacking. I have therefore conducted a study on the reliability of crowdsourced annotation for image annotations and their influence on system rankings in comparison to expert judgements. The results of this study are presented in Chapter 6.

4.3. Performance measures for image annotation evaluation

Multi-label classification evaluation often follows the evaluation per concept paradigm adopted from categorisation evaluation. This issue is further analysed by examining the usage of performance measures in related work on image annotation and by classifying them according to their

Ground Truth				System			
	Person	Blurry	Tree		Person	Blurry	Tree
Photo1	0	1	1	Photo1	0.1	0.9	0.1
Photo2	0	0	1	Photo2	0.2	0.9	0.7
Photo3	1	1	0	Photo3	0.6	0.2	0.8
...				...			
Photo1	0	3	2	Photo1	0	1	0
Photo2	0	1	4	Photo2	0	1	1
Photo3	2	1	0	Photo3	1	0	1
...				...			

Threshold

Figure 4.1.: Dimensions of evaluation measures

characteristics in the following.

The review of performance measures for VIR evaluation in Section 3.3 reveals three dimensions into which performance measures can be grouped (see also Figure 4.1):

1. Measurement direction:

Measures are classified according to the measurement direction. Horizontal measures evaluate performance per media item, while vertical measures consider performance per concept.

2. Prediction format:

Measures differ in the considered prediction format. Some work directly on ranked predictions of a system, while others consider thresholded, binary decisions for presence and absence of concepts. In the first case, the *cut-off* level is left to the effectiveness measure, while in the second case, the system itself accounts for the partition into relevant and irrelevant sets.

3. Relevance format:

Finally, measures can be distinguished concerning the relevance format that they support. Part of the measures take binary relevance of the ground truth into account, while others rely on graded relevance.

4.3.1. Performance measures in image annotation literature

Related work on image annotation was presented in Section 2.4. In the following, these works are analysed according to the performance measures that were adopted in the experiments. Working notes and summary papers of image annotation benchmarks, as well as review papers are excluded. The adopted measures in image annotation benchmarks are analysed separately. In total, 34 publications on image annotation are considered in this analysis. The results are depicted in Table 4.1. Approaches considering all characteristics of one dimension are counted in all three columns of the respective dimension.

A large amount of work focuses on concept-based evaluation. The evaluation with binary predictions is somewhat more popular than the one with ranked predictions, but the data gives no clear preference for either characteristic of this dimension. A third of the reviewed approaches considers binary and ranked predictions simultaneously for evaluation. Only one approach relies on graded relevance for evaluation. Presumably, evaluation with graded relevance is not often employed in image annotation due to the fact that available test collections mostly do not support graded relevance judgements. Therefore, the relevance format dimension is not further taken into account in this thesis, and the following discussion on performance measures is fixed to binary

Table 4.1.: Number of approaches in state-of-the-art image annotation that consider different dimensions of performance measures. Works that take into account several performance measures in the evaluation are included in all of the respective columns.

	concept-based	example-based	both
# works	31	7	5
	binary prediction	ranked prediction	both
# works	24	18	10
	binary relevance	graded relevance	both
# works	33	1	0

Table 4.2.: Number of approaches that consider different characteristics of evaluation dimensions.

	concept-based	example-based
binary prediction	22	3
ranked prediction	18	2

Table 4.3.: Performance measures for multi-label annotation sorted according to the horizontal-vertical dimension and prediction dimension. Measures in bold have been applied for performance measurement in image annotation literature.

	concept-based measures	example-based measures
binary predictions	Precision Recall F-Measure Accuracy	Precision Recall F-Measure Accuracy Alpha Evaluation Hamming Loss
ranked predictions	AUC EER Interpolated AP Mean AP R-Precision Precision@k Precision-Recall curve	One Error Coverage Ranking Loss Mean AP R-Precision Precision@k ranked accuracy

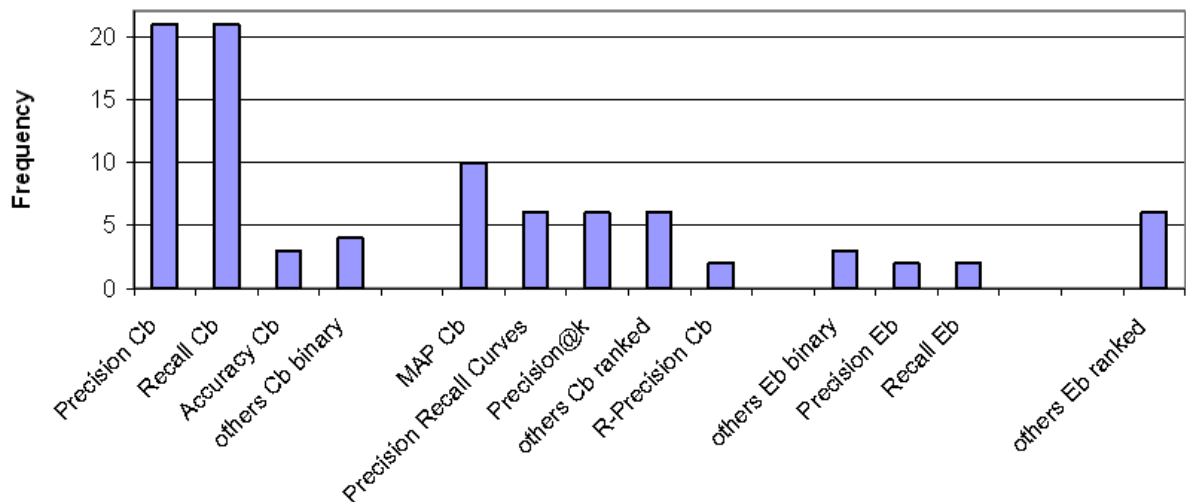


Figure 4.2.: Frequency of the usage of performance measures in image annotation literature.

relevance. Table 4.2 illustrates the number of approaches that adhere to characteristics of the remaining two dimensions for evaluation. Again, the trend to use concept-based evaluation is dominant, while the frequencies of using binary versus ranked predictions are almost equal.

Table 4.3 summarises the most important measures that were introduced in Section 3.3, structured by the direction and prediction format dimensions. Measures that were utilised in the reviewed literature on image annotation are highlighted in bold. The table shows that measures from all four characteristics of both dimensions are in use. Further, the diagram in Figure 4.2 demonstrates the frequency of the adopted measures in literature. Here, a clear dominance of concept-based Precision and Recall, followed by MAP, is apparent.

4.3.2. Performance measures in VIR benchmarks

Table 4.4 illustrates which measures were adopted in image and video annotation tasks of benchmarking campaigns. Similarly to related work, Precision, Recall, PR-curves, and variants of AP are often employed. Additionally, the benchmarks frequently report ROC curves, including derived measures such as EER and AUC, and the Error Rate. These measures are not as often reported in state-of-the-art literature. Please note that several benchmarks report more measures to the participants (i.e., they run the `trec_eval` software and provide the complete output). The listed measures refer to the ones used to rank systems and that are reported as main measures on the benchmark websites. TRECVID introduces pooling for the performance evaluation from 2006 on and employs the `infAP` measure. This is the only annotation task which uses the pooling strategy. All other tasks rely on a fully assessed test collection. Further, despite the ImageCLEF Photo Annotation task, all benchmarks evaluate in a concept-based fashion. The Photo Annotation task included example-based evaluation in 2009. Please note that I have been responsible for the organization of this task from 2009 onwards. The measures used in the Photo Annotation task denote the central point of this thesis and will be introduced in Chapter 7, 8, and 9. More information on challenges of the task and results from participants are reported in Chapter 10.

The analysis of the tasks until 2009 reveals that the reported measures all consider each concept in a binary fashion, despite the ImageCLEF Medical image annotation task starting in 2007 and the work of Mei et al. (2008). Correct predictions (TPs and TNs) score with 1, and false predictions (FNs and FPs) add 0 to the result score.⁴ However, some false predictions may be less wrong

⁴Some measures consider costs instead of scores and count correct and false predictions with 0 and 1.

Table 4.4.: Overview of performance measures utilised in image and video annotation benchmarks sorted by the year.

Benchmark & Task	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
TRECVID high-level feature extraction	precision, recall, PR-curves, AP	precision, recall, PR-curves, AP	precision, recall, PR-curves, AP	precision, recall, PR-curves, AP	precision, recall, PR-curves, infAP	precision, recall, PR-curves, infAP	PR-curves, infAP	PR-curves, infAP	xinfAP	xinfAP
ImageCLEF Medical Annotation	x	x	x	Error Rate	Error Rate	hierarchical error counting scheme	hierarchical error counting scheme	hierarchical error counting scheme	x	x
ImageCLEF Photo Annotation	x	x	x	x	Error Rate	EER, AUC, ROC	EER, AUC, ROC	EER, AUC, ROC, OS ^a	iAP, F _{ex} , OS-FCS ^b	iAP, F _{ex} , SR-Precision ^c
ImageCLEF Robot Vision	x	x	x	x	x	x	x	own scoring scheme	own scoring scheme	x
ImageCLEF Medical Modality Classification	x	x	x	x	x	x	x	x	accuracy	accuracy
PASCAL VOC classification	x	x	x	EER, AUC, ROC, PR-curves	EER, AUC, ROC, PR-curves	PR-curves, iAP	PR-curves, iAP	PR-curves, iAP	PR-curves, AP	PR-curves, AP
PASCAL VOC detection	x	x	x	iAP	iAP	PR-curves, iAP	PR-curves, iAP	PR-curves, iAP	PR-curves, AP	PR-curves, AP
VideoCLEF / MediaEval Tagging tasks	x	x	x	x	x	x	F-measure	MAP, mean reciprocal rank	MAP	MAP
ROMIP Image annotation	x	x	x	x	x	x	x	x	precision, recall, accuracy	no information
ImageNET challenge	x	x	x	x	x	x	x	x	flat+hierarchical error	flat+hierarchical error

^aSee the definition of this measure in Chapter 7^bSee the definition of this measure in Chapter 8^cSee the definition of this measure in Chapter 9

(i.e., predicting river instead of sea) than others (i.e., predicting cars instead of trees). I would like to add a fourth dimension for categorising evaluation measure, the *score prediction* dimension. This dimension differentiates if binary scores are assigned to correct and false predictions, or if the scores are based on a semantic relatedness between concepts.

4.3.3. Discussion

The preceding analysis reveals that performance measures for image annotation consider the measurement direction and prediction format dimensions. The relevance format dimension is rarely in use, as is the score prediction dimension. In this thesis, I would like to focus especially on the score prediction dimension, in addition to the measurement and prediction format dimensions. In particular, questions regarding a semantic embedding of the computation of performance scores will be answered. The ground truth of image annotation collections is assumed to exist in binary relevance format, which represents the common choice in multi-label annotation evaluation. In the following, a review of related work about performance measures that consider semantics is provided. Although these measures are not established as performance measures for image annotation, there is an active research field dealing with determining the degree of misclassification in IR systems. Often, these measures only work for the categorisation case and not for the multi-label annotation case.

4.4. Performance measures considering semantics

In literature, different performance measures have been introduced that consider the spatial and semantic relation between concepts in the score computation. A considerable amount of performance measures focuses on hierarchical evaluation. These rely on a structure of the concepts to determine a score. Some have been proposed to be applicable in the evaluation of hierarchical classifiers. In the following, different hierarchical evaluation measures that take the spatial relation between concepts into account are introduced. Most of them also consider the degree of misclassification to determine the score.

4.4.1. Hierarchical measures for uni-label classification evaluation

Hierarchical evaluation measures require a hierarchical structure of concepts to be applicable to an evaluation scenario. Different hierarchical measures for uni-label classification are summarised in Freitas and de Carvalho (2007). Intuitively, the concepts located close to each other in a hierarchy are more similar than the ones that are far apart. The idea is to judge an annotation from the predictor that does not exactly match the ground truth by their distance in the hierarchy. The most important measures are the depth-independent distance-based misclassification costs (DIMC) and the depth-dependent distance-based misclassification costs (DDMC). In the former case, the predicted concept is compared to the correct one and the number of edges of the shortest path in the hierarchy between both are counted. In the latter case, an additional weight is assigned to each edge in the hierarchy. Thus, misclassifications in deeper levels of the hierarchy are assigned lower costs than at an upper level (see also Blockeel et al. (2002)). Maynard et al. (2006) present the balanced distance metric to augment the standard precision and recall measures with a distance term for the false positive and false negative terms. The balanced distance metric takes the relative specificity of the taxonomic positions of the predicted and true label into account. Therefore, the measure is related to cost functions based on depth-dependent misclassification costs. A review of hierarchical performance measures for uni-label classification evaluation can also be found in Costa et al. (2007).

4.4.2. Hierarchical measures for multi-label classification evaluation

In the following, I focus on related work that proposes hierarchical evaluation measures applicable to multi-label classification evaluation. In multi-label classification evaluation, the way to deal with partly incorrect label sets must be defined. The above mentioned evaluation measures all base a cost or a similarity on the distance between two concepts in the associated hierarchy. In the case of multiple labels, a way must be found to determine which label from one set should be related to which label from the other set. Often, the predicted and the true label sets are compared concept by concept in a binary fashion. For each disagreement, the proposed evaluation formula is applied instead of considering the misclassification directly as the maximum error. An overview on hierarchical classification performance measures that are applicable to multi-label classification can also be found in Silla and Freitas (2010).

4.4.2.1. Hierarchical loss measures

In Cesa-Bianchi et al. (2006), a hierarchical loss function (H-Loss) that considers classification into a hierarchy with multiple and partial paths is proposed. The first wrongly classified node is regarded as a mistake and adds to the loss. Mistakes in the same subtree are not considered further based on the intuition that if a coarse classification fails, the fine-grained classification will also be incorrect, and is thus unimportant to evaluate. The underlying assumption is that for each classification a path from root to leaf or from root to an internal node is present. The smaller the value of H-Loss, the better the prediction of the system. The authors compare their work to the zero-one loss, which adds to the loss if the predicted set and the ground truth set of labels are not equal, and the Hamming loss (see Equation 3.18). In Cai and Hofmann (2007), another loss function for the evaluation of hierarchical multi-label classification is proposed. The so-called *maximal loss* extends the Hamming Loss (see Equation 3.18) by expanding the label sets with its ancestors in the hierarchy. Further, they propose the *parent one-accuracy*. This measure is an adaptation of the One-Error measure for hierarchical classification, as it calculates One-Accuracy at the parent node level of the concept. In contrast to One-Error, One-Accuracy is defined to add one to the score in case the highest ranked label belongs to the set of system labels (in contrast, One-Error adds one in case the highest ranked label does not belong to the set of system labels, see Chapter 3.3.5, (Cai (2008))). Struyf et al. (2005) propose a weighted Euclidean distance to compare two sets of labels. The weight is a concept-dependent factor that is dependent on the depth of the concept in the hierarchy. It is determined by $w_c = 0.75^{\text{depth}(c)}$.

4.4.2.2. Augmented versions of Precision, Recall and F-measure

Sun and Lim (2001) extend the standard Precision and Recall measures by an *Average Category Similarity* term and reformulate the standard F1-measure to include the category similarity. The category similarity between two categories is based on the cosine distance between the feature vectors of the particular categories. The average category similarity contains the contribution of the predicted label set to the ground truth label set for all missed and falsely classified labels. It is used to replace the strict FP and FN calculation in Precision and Recall, as well as in other set-based performance measures, and assigns fine-grained costs. A second proposal of the authors replaces the category similarity term with a category distance term. The distance is determined by the number of links in the hierarchy between both categories. Another hierarchical extension of the example-based Precision, Recall, and F-measure has been proposed by Kiritchenko et al. (2005). Their approach calculates the standard measures on the augmented predicted and ground truth label sets in which all concept ancestors of a label are additionally included. A similar modification

was proposed in Verspoor et al. (2006), while in Ipeirotis et al. (2001), an augmentation of the sets with the descendants in the hierarchy is proposed.

4.4.2.3. Hierarchical label-to-label distance functions

Wu et al. (2005) suggest a multi-label classification evaluation measure based on the shortest path in the hierarchy between labels. The problem of relating labels of the predicted set with labels in the ground truth set is solved by adding the dummy label 0 to the smaller set until the cardinalities of both sets match. The costs between two labels are determined by weighted costs according to the occurrence of the label in the training data, and by the position in the hierarchy. Then, the average distance of the labels between both sets is calculated by minimising the costs of the bijective mapping.

4.4.3. Ontology-based performance measures

Shevade and Sundaram (2006) put forward an *average ConceptNet similarity* measure. It is based on ConceptNet, which is an ontology of common-sense knowledge that supports 20 different semantic relations between concepts. The distance between two synsets of ConceptNet is determined by the Hausdorff distance between both sets. It is converted into a similarity and averaged over all media items in order to obtain the average ConceptNet similarity. Cordi et al. (2005) propose to base the similarity computation between concepts on an extension of the Dijkstra algorithm applied to the nodes of an ontology. Pairs of concepts of the ontology are supposed to have weights assigned. These weights are manually assigned by the ontology designer and denote the similarity between concepts. The similarity is iteratively computed between all predicted concepts and the set of target concepts, and is later averaged. Mei et al. (2008) propose the Relevance Comparative Score. It is based on a semantic distance of label sets in WordNet by employing the JCN semantic relatedness measures (see Section 4.5 for a discussion on semantic relatedness measures). The performance measure works only in the comparison of two approaches, as it assigns scores relatively to each other. The resulting effectiveness score cannot be related to the effectiveness score of other comparisons. This performance measure was directly employed for multi-label image annotation, but has not been used in other research or accepted by the research community so far.

4.4.4. Comparison

Table 4.5 shows a classification of the hierarchical performance measures for multi-label classification evaluation. The comparison focuses on the granularity of costs assignment (equal costs from an hierarchy, adapted costs based on the depth in the hierarchy, and other ways to obtain costs) and the restrictions in different classification evaluation scenarios (uni-label hierarchical classification, multi-label hierarchical classification, and free multi-label classification). Please note that hierarchical approaches which can deal with multi-label classification evaluation are also applicable to uni-label classification most of the time. Some hierarchical performance measures are constrained to a tree-based structure of the graph and cannot deal with directed acyclic graphs. The measures listed under free multi-label classification make no assumptions about the topology of the graph.

4.4.5. Discussion

This section reviews performance measures that include semantics to calculate an evaluation score. Most rely on an external hierarchy to determine the costs between misclassified concepts, while

Table 4.5.: Comparison of hierarchical classification approaches.

Approach	uniform hierarchical costs	adaptive hierarchical costs	other costs	hierarchical uni-label classification	hierarchical multi-label classification	free multi-label classification
depth independent misclassification costs (Blokkeel et al. (2002))	X			X		
depth dependent misclassification costs (Blokkeel et al. (2002))		X		X		
balanced distance metric (Maynard et al. (2006))		X		X		
h-loss (Cesa-Bianchi et al. (2006))		X			X	
precision and recall with category distance (Sun and Lim (2001))	X				X	
maximal loss (Cai and Hofmann (2007))	X				X	
weighted Euclidean distance (Struyf et al. (2005))		X			X	
hierarchical precision, recall, f-measure (Kiritchenko et al. (2005))	X				X	
hierarchical precision, recall, f-measure (Verspoor et al. (2006))	X				X	
hierarchical precision, recall, f-measure (Ipeirotis et al. (2001))	X				X	
shortest path between label sets (Wu et al. (2005))		X			X	
extension of Dijkstra (Cordi et al. (2005))		X			X	
precision and recall with category similarity (Sun and Lim (2001))			X			X
average ConceptNet similarity (Shevade and Sundaram (2006))			X			X

some use ontologies. Except Mei et al. (2008) and Shevade and Sundaram (2006), the original publications of these measures do not consider image annotation as an application domain. They are mainly focused on text classification and gene prediction. A detailed discussion on the strength and limitations of these measures and a proposal of a new hierarchical performance measure for multi-label image annotation evaluation are presented in Chapter 7.

4.5. Semantic relatedness measures

Two terms are semantically related if they form a certain relationship such as an *is-a* relation, *part-of* relation (meronym) or antonym relation (words with opposite meanings). The term “semantic relatedness” is therefore a generalisation of the term “semantic similarity,” as it includes several kind of relations (Resnik (1999)). Ways of assessing the semantic relatedness between two terms have gained a great deal of attention from researchers for several years. One simple approach analyses the co-occurrence of concepts in the training set and uses them as a semantic indicator of which concepts are likely to appear simultaneously and are therefore semantically related. While this approach is easy to set up, it is directly dependent on the test collection of an image annotation system and does not explicitly and generally model relationships between concepts. It is heavily biased towards the topics that occur in the collection and strongly dependent on its coverage. This dependency on the utilised corpus is also analysed in Lindsey et al. (2007).

This section introduces several semantic relatedness measures that are not directly bound to an image annotation test collection, but rather consider other, more general, information sources. The semantic relatedness measures are differentiated according to the information source from which the relatedness is determined: Thesaurus-based, Wikipedia-based, and Web-based information sources. These measures were not originally proposed for performance evaluation, but determine general ways of estimating the semantic relationship between words. However, they carry the potential to be used as indicators for the degree of misclassification in performance measurement. Please note that part of this section has been published in Nowak et al. (2010) and includes joint work with Ainhoa Llorente from the Knowledge Media Institute, The Open University, Milton Keynes, UK.

4.5.1. Thesaurus-based approaches

Semantic relatedness measures based on a thesaurus have a long history in natural language processing research. Thesaurus-based methods rely on a hierarchical representation of concepts and relations as nodes and links, respectively. A fair amount of thesaurus-based semantic relatedness measures were proposed and investigated on the WordNet hierarchy of nouns (see Budanitsky and Hirst (2006) for a detailed review). They can be differentiated into path length measures, information content measures, and gloss-based measures.

Path length measures consider the distance between concepts in the WordNet hierarchy. Much like the hierarchical measures introduced in Section 4.4.1, they assume that similar concepts are located in close proximity to each other. Wu and Palmer’s measure (WUP) (Wu and Palmer (1994)) defines the conceptual similarity between two terms in a hierarchy by the path length between these terms to their least common super-concept (LCS) and the length between the LCS and the root. Leacock and Chodorow’s measure (LCH) (Leacock and Chodorow (1998)) investigates the path length between each sense of two concepts in WordNet and chooses the senses with the shortest path to define the semantic relatedness. The measure PATH is equal to the inverse of the shortest path length between two terms. Hirst and St-Onge’s measure (HSO) (Hirst and St-Onge (1998)) is based on lexical chains of semantically related terms. The relatedness depends on the length of the chain and on the number of changes in direction.

Path length measures assume that links in the ontology represent uniform distances. Although the WordNet ontology was constructed with care, this is not always the case, as is for example seen in the density degree in subtrees. The IC is not sensitive to varying link distances, as it denotes the frequency of occurrence in a large corpus. The IC of a concept is quantified as its negative log likelihood. Intuitively, the more abstract a concept is, the lower the IC it conveys. Resnik's measure (RES) (Resnik (1995)) introduces a semantic similarity in a taxonomy which is based on the amount of IC two terms share. Jiang and Conrath's measure (JCN) (Jiang and Conrath (1997)) combines the RES measure with the shortest path measure by using the IC as decision factor. Lin's measure (LIN) (Lin (1998)) scales the IC of the most specific concept that subsumes both concepts by the sum of the IC of the individual concepts. The main disadvantage of information content measures lies in the frequency calculation. Its generalisation capabilities are strongly depended on the coverage of the corpus used.

Gloss-based measures work on the short textual definition of concepts, which is called *gloss* in WordNet. Lesk's measure (LESK) was introduced in Lesk (1986) and later adapted by Banerjee and Pedersen (2003). It calculates the relatedness of terms based on the extent of overlaps in their WordNet definition by utilising the gloss overlap measure. Patwardhan's measure (VEC) (Patwardhan (2003)) uses a co-occurrence matrix from a corpus made of WordNet glosses and associated context vectors. The relatedness is measured by calculating the cosine between corresponding context vectors. The main drawback of gloss-based measures occurs in cases when there exists no shared word between two glosses.

4.5.2. Wikipedia-based approaches

Further research use the Wikipedia corpus to determine semantic relatedness. Strube and Ponzetto (2006) test several classic WordNet measures on the Wikipedia corpus. Gabrilovich and Markovitch (2007) propose Explicit Semantic Analysis on the Wikipedia corpus which represents the meaning of words in a high-dimensional space derived by a machine learning procedure. Milne and Witten's measure (WIKI) (Milne and Witten (2008)) calculates the relatedness of two terms by using the angle between vectors of associated links in the hyperlink structure.

Wikipedia benefits from its hyperlink structure. Compared to WordNet, the hyperlinks offer more flexible relationships than the *is-a* relationship used in WordNet. Wikipedia-based measures as well as WordNet-based measures need a disambiguation part in which the word is assigned to the searched sense. Zesch and Gurevych (2010) present a comprehensive study of different WordNet and Wikipedia measures and evaluate them on different datasets with human similarity judgements. They conclude that the Wikipedia corpus does not generally outperform the WordNet thesaurus (opposed to earlier research), as most differences are not significant.

4.5.3. Web-based approaches

More recently, several semantic relatedness measures based on search engine results have been proposed. Often, they are referred to as *distributional methods*, as they define the semantic relatedness between two terms as their co-occurrence in similar contexts. In contrast to the WordNet measures, their strength lies in a maximum coverage, domain independence, and universality (Gracia and Mena (2008)). A differentiation is made between distributional methods that rely on text documents to extract the knowledge and distributional methods that gain the knowledge from images and associated metadata. The former are called methods relying on document-based information, the latter methods based on image resources.

Document-based measures: Recent approaches use the World Wide Web (WWW) as a corpus for distributional semantic relatedness estimation. Some works employ text snippets, as shown in Sahami and Heilman (2006), or use text snippets in combination with page counts, such as in Bollegala et al. (2007). However, they are limited to semantic similarity and do not generalise to semantic relatedness. Haubold and Natsev (2008) propose to determine the IC that is employed in several information content approaches of WordNet measures from a collection of Web documents instead of the Brown corpus. They show that the better coverage of terms in Webpages allows a better estimation of the IC. Other approaches determine the correlation between terms by crawling them with Web-search engines and by weighting the results based on some distance criterion. Cilibrasi and Vitanyi (2007) propose the Normalised Google Distance (NGD) and generalise it to the Normalised Web Distance (NWD) depicted in Equation 4.1.

$$\text{NWD}(x, y) = \frac{\max\{\log h(x), \log h(y)\} - \log h(x, y)}{\log N - \min\{\log h(x), \log h(y)\}}, \quad (4.1)$$

where $h(x)$ denotes the number of documents in which concept x occurs, and $h(x, y)$ denotes the number of documents associated with both concepts x and y . The parameter N refers to the total number of documents available. Gracia and Mena (2008) extend the NWD by applying a transformation to obtain a relatedness measure in the range of $[0; 1]$ which increases with decreasing distance.

$$\text{rel}(x, y) = e^{-2\text{NWD}(x, y)}. \quad (4.2)$$

Document-based measures need no disambiguation task, have a better word coverage and do not have to deal with missing terms. Relationships between words are not limited to the *is-a* relation as in WordNet. However, it is not clear whether textual resources are the correct means to estimate correlations for visual concepts.

Image-based measures: The document-based distributional measures rely on the number of hits in textual documents retrieved by Web search engines. It is rather questionable whether these textual documents can represent the co-occurrence relationship of visual concepts adequately. Consequently, research investigated the utilisation of information from photo communities such as Flickr for the definition of semantic relatedness between concepts. Jiang et al. (2009) propose the Flickr Context Similarity (FCS). They use the Flickr search functionality to search for concepts in image tags, descriptions, and comments, and apply Equation 4.2 to estimate a relatedness value. In their work, they utilise the FCS to automatically select video concept detectors from a pool of detectors to answer a user query. Additionally, they perform a small experiment between the Flickr Tag Similarity (FTS) (just searching the Flickr tags) and the FCS, and conclude that FCS has a better word coverage. Wu et al. (2008) propose the so-called Flickr distance which quantifies semantic relationships between concepts in the visual domain using a machine learning approach. For each concept, 1000 images are downloaded, visual features are extracted, and a latent topic visual language model is computed. Finally, they define the Flickr distance between two concepts as the average square root of the Jensen-Shannon divergence between the two latent topic visual language models associated to them. This method is computationally expensive and relies on visual features as well as on the language model as adjustable parameters. Liu et al. (2007) propose a linear combination of a statistical and a visual correlation term. They apply a variation of NGD to Google Image Search arguing that an image search engine has a better coverage of visual concepts than a text search engine to find word correlations. The second correlation term is determined by the visual consistency of the top 20 retrieved images.

The image-based measures benefit from the same advantages as the document-based measures, which includes a wide coverage, no disambiguation and no missing terms. Additionally, image

resources might provide a better corpus for semantic relatedness estimation in the application for image annotation and evaluation.

4.5.4. Discussion

This section presents research on the determination of semantic relatedness between words. All approaches propose an automated method to derive similarities which can be used in order to enhance annotation and retrieval systems. In the area of image annotation, several semantic relatedness measures have been investigated in order to improve the classifier output, such as in Jin et al. (2005), Liu et al. (2007), Llorente et al. (2009), or Qi et al. (2009). To my knowledge, none of these measures has been explicitly used for performance evaluation except the JCN measure used by Mei et al. (2008), as presented earlier. Despite some overlaps between the hierarchical performance measures, such as DIMC and the WordNet based measures, this branch of research has been used in other scenarios. Furthermore, the evaluation of annotation approaches requires a mapping of a set of ground truth labels to a set of predicted labels. The semantic relatedness measures do not provide any information on how this mapping can be performed in an evaluation scenario, and on which words should be related to each other. A performance measure which considers semantic relatedness measures to determine the degree of misclassification will be presented in Chapter 8. Several semantic relatedness measures, including WordNet measures, Wikipedia measures, and distributional measures on textual and visual corpora, will be investigated concerning their applicability in image annotation evaluation. Moreover, results on the correlation of human similarity judgements to the relatedness scores will be provided and compared with the state-of-the-art.

4.6. Evaluating evaluation measures

Although experimental evaluation in IR and VIR follows an accepted methodology when running retrieval experiments, the agreement on which performance measure is best remains an issue of discussion. Often, the question about the desired outcome of an evaluation measure is neither easy to define nor to prove. As Dupret and Piwowarski (2010) tellingly describe: “*Deciding which metric is best calls for a third ‘meta’ metric. Because various ‘meta’ metrics are likely to co-exist, a meta metric for the meta metrics is necessary, etc.*”

On one hand, a performance measure should relate to the user model and evaluation objective as detailed in Section 3.3, because experimental evaluation in IR is an abstraction of a user’s search. It should therefore show a high correlation with human judgements. On the other hand, necessary and desired requirements on metrics can be objectively defined as in King (2003):

1. A metric should reach and only reach its highest value with perfect quality.
2. A metric should reach its lowest value only for the worst possible quality.
3. A metric should be monotonic in the sense that a better quality goes with a higher score.
4. A metric should be clear and intuitive.
5. A metric should correlate well with human judgements.
6. A metric should be reliable in the sense that it exhibits little variance for equivalent inputs.
7. A metric should be cheap to set up and apply.
8. A metric should be automated.

9. A metric should measure what it is supposed to measure.

Additionally, a metric is mathematically defined as a positive definite function (i.e., it is non-negative and preserves identity of indiscernibles (strictness)), to be symmetric, and to satisfy the triangle inequality. It is a pseudo-metric in case the strictness requirement is not fulfilled. Debates about the necessity of the symmetric property (i.e., researchers argue that similarity can also be asymmetric) and the triangle inequality property exist. A detailed discussion on this can be found in D'Amato (2007).

Despite these requirements, the effectiveness of one performance measure in contrast to another is one of the most important characteristics in meta-evaluation. Often, the benefit of measures is discussed based on their principles, sometimes based on empirical studies.

4.6.1. Rank correlation

Meta-evaluation incorporates empirical studies to assess the effectiveness of performance measures. Usually, correlation and stability experiments are performed which show the relative benefit of one measure to another (Sanderson (2010)). Tague-Sutcliffe and Blustein (1995) have established the methodology to take submitted runs of an evaluation benchmark in order to conduct these experiments without bias by a particular system. All runs are evaluated with different performance measures on a given test collection. The resulting effectiveness scores are ordered into a ranked list. Different performance measures produce different system rankings, and the goal of rank correlation analysis is to assess the degree of association between any two measures.

The most frequently adopted rank correlation measures, Kendall's τ , Kolmogorov-Smirnov's D , Spearman's ρ , and Pearson's correlation coefficient r are introduced in the following. The Kendall τ coefficient (Kendall (1938)) is a non-parametric statistic that is used to measure the degree of correspondence between two rankings and to assess the significance of this correspondence. It counts the minimum number of pairwise adjacent swaps by turning one ranking into the other.

$$\tau = \frac{P - Q}{P + Q}, \quad (4.3)$$

with P being the concordance count and Q the discordance count. After normalising, two identical rankings produce a correlation of +1, the correlation of inverse rankings is -1, and the expected correlation of two rankings chosen randomly is 0. The Kendall τ statistic assumes a discordant ranking as null hypothesis and rejects the null hypothesis when τ is greater than the $1 - \alpha$ quantile, with α as significance level.

Kolmogorov-Smirnov's D (Kolmogoroff (1933)) also uses a non-parametric statistic, but it states as null hypothesis that the two rankings are concordant. In Kolmogorov-Smirnov's test, two rankings are compared by means of the "uniform" distribution function F induced by \mathbf{x} and the empirical distribution function F induced by \mathbf{y} , that is, by calculating the number of objects not following i in \mathbf{y} for every i .

$$D = \max_i |F_x(i) - F_y(i)|. \quad (4.4)$$

Spearman's ρ rank order correlation coefficient uses the ranks in the system rankings to sum up the distances of disorders (Gibbons and Chakraborti (2003)):

$$\rho = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (S_i - \bar{S})^2}}, \quad (4.5)$$

with R_i denoting the rank of effectiveness score x_i and \bar{R} as mean rank for the ordered list x . S_i and \bar{S} are defined analogously for the ordered list y .

Pearson's linear correlation coefficient r is closely related to Spearman's ρ . It does not consider the ranks, but rather works directly on the effectiveness scores that were assigned by the performance measures. Pearson's correlation coefficient is defined as:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (4.6)$$

with $\bar{x}_i = x_i - \bar{x}$ indicating the difference of the score x_i to the mean effectiveness score of list x and $\bar{y}_i = y_i - \bar{y}$ denoting the difference of the score y_i to the mean effectiveness score of list y .

Several statistics have been criticised in the community due to different limitations. Melucci (2007) illustrates that it is likely that the Kendall τ statistic rejects the null hypothesis and decides for concordance, e.g., if the sample size is large. In his work, he compares the τ statistic with the Kolmogorov-Smirnov D statistic and recommends using several test statistics to support or revise a decision. According to Melucci (2007), the Kolmogorov-Smirnov D statistic is less affected by the sample size, is sensitive to the extent of disorder (in contrast to τ , which takes the number of exchanges into account), and tends to decide for discordance, for instance, in the case of two radically different retrieval algorithms. Further, Carterette (2009) criticises Kendall's τ because it treats all pairwise swaps equally, assumes statistical independence, and shows a high variance over the system sample space. He proposes the rank distance measure which accounts for the problems with Kendall's τ , although providing an unintuitive measure. Yilmaz et al. (2008) propose AP correlation to account for the equal penalisations of rank disorders in Kendall's τ throughout the entire ranking. Kumar and Vassilvitskii (2010) extend Kendall's τ with element weights, position weights, and pairwise distances between permutations. Zesch and Gurevych (2010) compare Spearman's ρ and Pearson's r . They show that Pearson's r is sensitive to outliers and that it measures the linear relationship between both rankings. In case the actual relationship is not linear, the results are flawed. Further, Pearson's r requires a normal distribution and interval scaled data. In contrast, Spearman's ρ is robust against outliers and can measure the association of non-linear relationships without assumptions on interval scales and on the distribution of random variables. But it is known that Spearman's ρ tends to give higher values for datasets with many tied ranks.

Despite the criticism, Kendall's τ is used as the standard statistic to compare the correlation of two ranked lists, while Pearson's r is adapted as the standard in measuring correlations of semantic relatedness measures to human judgements. Sakai (2006) compares the Kendall's τ rank correlation with bootstrap estimates of standard error for the rank correlations in order to quantify the accuracy of rank correlation between two IR metrics, and concludes that the rank correlation estimates are quite accurate. Thom and Scholer (2007) investigate the correlation of different performance measures in search tasks adopting Kendall's τ . Cormack and Lynam (2007) use significance tests as the t-test, Wilcoxon test, and sign test to compare two evaluation measures, and conclude that especially the t-test shows superior qualities. Moffat and Zobel (2008) use Kendall's τ as well as significance tests to analyse properties of different measures in text retrieval tasks in addition to an analysis of desired properties according to a user model. Kendall's τ is further employed in Sakai (2008) and Sanderson and Joho (2004), while Webber et al. (2007) use Pearson's and Kendall's rank correlation.

4.6.2. Stability analysis

The stability of a performance measure is the second property most measures are compared against. Sometimes, it is also denoted with discriminancy, as it determines the ability of a measure to discriminate between different systems on several test collections.

Zobel (1998) investigates the reliability and stability of retrieval results with incomplete relevance judgements. Buckley and Voorhees (2000) compute error rates on retrieval measures by

varying the expression of 50 topics in 21 versions. In their experiment, the change of the expression did not change the query topic and therefore kept the same information need. As a result, the number of relevant documents per topic remains stable, while retrieval systems produce changes in the retrieved document sets; this allows the computation of confidence intervals. They find that there exists a large inverse correlation between the number of documents considered by a performance measures (i.e., Precision@10 vs. Recall@1000) and the size of the error rate. Further, the stability of Precision@k measures with cut-off values for $k < 30$ is low in contrast to measures like MAP, R-Precision, or Precision@1000. In subsequent work, Voorhees and Buckley (2002) investigate the probability of one effectiveness measure deciding that system A is superior to system B on one collection, while it concludes the opposite on another collection. They investigate if the same run of a pair of runs still wins on two disjunct topic sets evaluated with MAP. Sanderson and Zobel (2005) look at the problem of effectiveness measure stability from the point of assessor costs. They conclude that Precision@10 is far more stable in comparison to MAP per equal quantity of assessor effort, as it requires only 10-14% of assessor effort. Aslam et al. (2005) propose a maximum entropy method to analyse performance measures. Sakai (2007) uses Voorhees' and Buckley's experimental design and compare graded relevance measures with binary relevance measures concerning their stability, sensitivity, and resemblance in ranking. Robertson et al. (2010) follow Sakai (2006) and Aslam et al. (2005) in evaluating their graded average precision measure based on its informativeness and its discriminancy. Further work on stability can also be found in Cormack and Lynam (2006), and in Sakai (2006).

4.6.3. Discussion

This section introduces requirements on performance measures and examines how the effectiveness of performance measures can be assessed. While there is ongoing research on effectiveness statistics, rank correlation coefficients, and stability experiments have proven to give valuable information on measure effectiveness despite all critics. Experiments on the effectiveness of evaluation measures conducted in Chapters 6, 7, 8, and 9 follow the outlined state-of-the-art on meta-evaluation. A detailed discussion on whether the established rank correlation and stability measures capture all desired properties for meta-evaluation is out of the scope of this thesis. Nonetheless, additional requirements on performance measures that directly relate to the use of semantics in image annotation evaluation will be identified. These requirements will be introduced according to a user model and include criteria such as random numbers or overannotation.

4.7. Summary

This chapter gives a thorough insight on how experimental evaluation is conducted and which means are used to assess the quality of retrieval and annotation systems. A special focus is placed on research that explicitly addresses evaluation methodologies for visual annotation approaches. Open issues are raised, and in cases where no works dealing with image or video annotation evaluation are available, methodologies from text-retrieval evaluation are detailed. While much experimental work has been conducted on the TREC collections, methodologies are reused in VIR and annotation evaluation without questioning or studying if the same assumptions hold. Although the conducted experiments have greatly influenced the progress in IR research and evaluation design, not all results should be taken for granted to hold in image annotation evaluation. It is not the goal of this thesis to question every result that was generated in over 50 years of text-retrieval benchmarking, but for selected sub-topics, experiments are adapted to image annotation needs and verified for visual collections. The biggest hypothesis in this work states that the evaluation

of image annotation poses its own demands on test methodologies and evaluation design. These demands will be analysed and experimentally proven in the following chapters.

The works reported in this thesis adhere to several technical choices. First of all, the experiments use the MIR Flickr dataset (Huiskes and Lew (2008)), a collection of 25,000 images including EXIF tags and Flickr user tags, published under the Creative Commons license. The collection satisfies the requirements addressed in Section 3.2.1. The relevance assessments of the original MIR Flickr collection are not used. Instead, the concept definition and relevance assessment process will be described in Chapter 5 and 6, respectively. In this thesis, experiments rely on the classic annotation task without any indication on the corresponding image region or segmentation masks. Further, the proposed performance measures for image annotation mainly address the score prediction dimension, while ignoring the relevance format dimension in the experiments due to the reasons detailed in Section 4.3.

Part II.

Test collections for multi-label image annotation evaluation

Outline

This part of the thesis details the contributions on test collections for image annotation. My work on visual test collections considers two parts: the definition of visual concepts according to user needs for image indexing and retrieval, and the assessment of relevance adopting crowdsourcing approaches. The outcome of a user study on how users prefer to structure their private photo collection is presented. Users are asked to split a set of photos hierarchically with a variation of the natural grouping approach and to assign keywords to photos in a free tagging task. A visual concept lexicon is defined based on the results of the user study, and the concepts are organised in the *Photo Tagging Ontology*. The second part deals with the costly process of relevance assessment. Recently, crowdsourcing approaches have been investigated in order to outsource judgement on relevance. This work analyses whether the ground truth acquisition in image annotation can be outsourced and investigates the reliability of crowdsourced assessments. Effects on system ranking in benchmark-like evaluation and inter-annotator agreement among experts and non-experts are determined. The results establish crowdsourcing as a fast and cheap alternative to expert judgements while retaining sufficient reliability. The exploitation of user-generated tags for the automated construction of training sets is a second option to obtain cheap annotations. Benefits in applying a one-class SVM in order to separate relevant from irrelevant images for a concept will be detailed.

relevance judgements. This chapter focuses on the definition of topics and their structuring in the form of an ontology. Topics, or, in this case, visual concepts, are strongly related to user needs and define the use cases of a system. While concepts are modality-independent (i.e., the event “birthday” might be detectable in the visual modality (birthday cake, people celebrating) as well as in the auditory modality (people singing a birthday song)), visual concepts are solely described by the visual content of a photo and are therefore language independent. The usage of visual concepts in common visual test collections was analysed in Chapter 4.1. A variety of the number of concepts as well as the structuring of concepts could be found. In the following, the focus is placed on user studies on organising photo collections and on the issue of whether visual concepts can be verified and derived from the results.

First, let me introduce some technical choices for test collections in this work. Throughout the thesis, the images of the MIR Flickr dataset (Huiskes and Lew (2008)), respectively parts of the set, are used as documents in the test collection. The MIR Flickr collection also defines a small set of visual concepts and provides relevance assessments. However, these concepts and assessments are not used for the experiments in this thesis. To better differentiate between the original MIR Flickr collection and the test collection consisting of the MIR Flickr images and metadata, but using other concepts and relevance judgements, the latter is referred to as *ImageCLEF Visual Concept Detection and Annotation Task (VCDT) collection*. Concepts are restricted to “visual concepts,” which means that the concept can be detected by using solely the visual content, and described by English terms. There are large taxonomies used for bibliographic indexing, e.g., in the Library of Congress, such as the Thesaurus of Graphical Materials with 69,000 subject terms alone (Hauptmann et al. (2008)). These taxonomies do not focus on visual concepts and do not solely contain terms that are automatically indexable. A review on these kind of taxonomies is out of the scope of this thesis.

5.2. User studies on photo organisation

Early work focuses on the use, requirements, and description of images by professional users. Markkula and Sormunen (1998) analyse user needs of journalists and group them into search categories of specific objects (named persons, building etc.), themes or abstractions from the photo, background information on an image, and known photographs. Armitage and Enser (1997) address and compare user needs for library queries in image archives. A review on early work about these specialised studies can be found in Eakins and Graham (1999).

A categorisation of image descriptions has been performed by Jörgensen (1998). She divides image attributes into 12 categories. The attributes were obtained in three user tests performing a viewing, a search, and a memory task on six images with a total of 107 participants. The categories include *Objects, People, Colour, Visual Elements, Location, Description, People-related attributes, Abstract concepts, Content/Story, External Relation, Viewer Response, and Art-historical information*. Greisdorf and O Connor (2002) provide users with terms belonging to one of the seven categories *Colour, Shape, Texture, Object, Action, Location, and Affect*, and ask them on one hand to use these pre-selected terms to describe images, and on the other hand to use free terms to associate them with images. They then analyse preferences for terms of special categories and conclude that affective and emotional terms are an important descriptive category in image retrieval, accounting for 39% of the user-supplied terms. These results rely on 10 images and the descriptions of 19 subjects. Hollink et al. (2004) ask subjects to read three text fragments and describe an image which illustrates each text fragment. Subsequently, the subjects searched for a fitting image using a Web image search engine. The classification of the queries and descriptions show that most terms belong to the conceptual level which includes objects and scenes. Laine-Hernandez and Westman (2006) build on the scheme proposed by Jörgensen (1998) after

reviewing the results of several studies in related work. They perform their own categorisation, keyword, and free description experiments on a broader test setup with 40 reportage-type images and 20 subjects. They suggest adding the three categories *Animal*, *Visual Quality*, and *Weather* to Jørgensen's scheme, as these categories were missing to adequately structure the obtained image attributes. Further, they suggest removing the category *Art-historical information*, as it finds no application in their experiment. Results show that most attributes belong to general, specific, and abstract semantic concepts.

Shifting to the issue of personal photo management, Rodden and Wood (2003) present one of the earliest studies on how people organise their own digital photo collections and compare it to the organisation of printed photos. They discovered that almost all participants organised their prints by putting them into albums, usually leaving out the "bad" photos which are of low quality or boring. Further, they analyse requirements on search in photo repositories. Sorted by decreasing frequency, these requirements were mentioned: searching for a particular event (e.g., a holiday), searching for an individual photo, and searching for a set of photos taken at different events that share a property such as depicting the same person. Further, people requested possibilities for the combination of characteristics, such as searching for a high quality photo of a special person. Kirk et al. (2006) draw implications for photo search and organising tools derived from the results of a user study. The time spent by users on sorting tasks underlines the central requirement on intelligent search tools. Like Rodden and Wood (2003), they name the opportunity for automated approaches to differentiate poor quality photos from the collection. Further, they conclude that search tools looking for objects, presence of people, or similar layouts are beneficial and allow for a different view on the photo collection. Miller and Edwards (2007) perform a study on the sharing of digital photos. As a by-product, they ask their participants about the organisation and tagging of photos and found that most participants are able to locate a special photo of their own collection just by a chronological sorting. However, many reported tagging as useful for others and that they assign tags such as "night" or "tree" to help other people find shots of desired objects.

In summary, the results of the different user studies reveal the following implications. Quality aspects represent a central requirement in which image annotation approaches should support the user. Mainly, the indexing of quality concepts serves the higher need for a fast rejection of poor images. Quality aspects are further not restricted to technical aspects of quality, but also include subjective opinions such as boring photos. People often use affective and abstract terms to refer to photos. An automated approach that can at least partially support these more subjective concepts, would be of help for the search of adequate photos. A central requirement for indexing personal photo collections is the detection and recognition of depicted people. While person recognition is a special field of image annotation and is not supported by the databases analysed and used in this thesis, the detection of persons as well as some personal characteristics, such as gender or age, denote a first step in realising user requirements. Next, events play a special role in the organisation of photo collections. Often, the photos are stored event-based on the hard disk and users often remember stories based on the event. Additionally, content-based elements of the photos, such as objects and scenes, should be indexed. While the results show that the photographer himself does not need this information to find a special photo, the benefits are apparent when photos are shared as the indexing provides better access for other users.

5.3. How do people describe and categorise images?

The user studies outlined above give a good insight into which user requirements should be addressed by photo management software and search engines. However, they do not answer the question of which content-based concepts are of interest in indexing. While the definition of attributes for describing images has been addressed in earlier studies (Jørgensen (1998), Greisdorf

and O Connor (2002), and Hollink et al. (2004), Laine-Hernandez and Westman (2006)), these studies rely on small datasets or specialised photos. In a user study on which concepts users prefer to describe and structure personal photos, we investigate if the results also hold for personal photo collections and whether they can be verified for a larger number of images. The user study was performed by one of my master thesis students and presented in Sieweke (2010). He uses two tasks, a natural grouping task and a free tagging task, to categorise photo descriptions and structure the descriptions in a hierarchy.

Natural grouping (Kuylen and Verhallen (1988)) is a method that asks subjects to successively split up a set of images into subsets. Each time a split is performed, the subjects describe why they chose to group part of the images in one set and the other part in the other set. This reveals information on which images are close to each other and which qualifications are chosen to create subgroups. In the original approach, reasons to split images into subgroups can only be used once in the whole process and different subsets cannot be combined again. Therefore, the first decision has a major influence on the whole hierarchy. We propose a modified version of the natural grouping approach, in which sets of different branches can be sorted in an alternative way after the second split. The tagging task asks the participants to freely associate attributes with images and therefore generates a view on which properties are used to describe images. In the study, participants were asked to associate five attributes to each image in order of importance. The user study considers 60 randomly sampled images from the ImageCLEF VCDT collection and 10 persons (5 male and 5 female) aged between 27-69 which participated in two sessions each.

The results of the natural grouping task show three decision categories. The first group considers object- or content-oriented decisions that are directly related to visual attributes of the image. Many of these decisions are related to persons (gender, age, body parts), daytime, special objects, and colour. The second group contains representational decisions focusing on the format of the image (i.e., close-up) and on the technical quality (i.e., blurry). The last group contains decisions based on affect and emotions such as bizarre or aesthetic. In total, 70% of the decisions was performed content-based, 13% were based on representational characteristics and 13% related to affective properties. The first split mostly differentiates between people and no people, respectively animate vs. inanimate. Two persons used the interestingness as first decision.

The results of the tagging task support the results of the natural grouping task. Again, attributes related to the content, representation, and affective characteristics of images were assigned. These groups can be subdivided into people (22.2%), nature concepts (16.6%), affective concepts (11.2%), urban concepts (11%), free associations (9.4%), animal concepts (5.3%), colour characteristics (4.1%), temporal characteristics (3.6%), events and activities (3.4%), format characteristics (3.3%), lightning effects (1.8%), and the assignment of no attribute (8.1%).

These results are in the line of research described in Section 5.2. Moreover, the attributes are directly assessed on a subset of the images which are used in the VCDT test collection. This implies that the distribution of attributes is likely to hold for the whole image set, while comparison to related work shows that the concepts can also be generalised with respect to other sets.

5.4. User tag analysis

Visual concepts in VIR are loosely related to user-generated tags. Web-based photo sharing sites allow the tagging of photos during or after upload. These tags are usually unconstrained, contain random noise, are imprecise, personalised, and limited in comparison to the search possibilities that would be desirable. In particular, they include information that might not be relevant to describe and retrieve images based on visual content, but are used for other purposes. On the other hand, tagging allows associating images with terms freely and unrestrictedly. The tags do not have to fit into a predetermined taxonomy and no problems with evolving tag hierarchies arise. Gupta et al.

(2010) define 11 kinds of user-generated tags which include content-based tags, context-based tags, attribute tags, ownership tags, subjective tags, or organisational tags, amongst others.

The motivation of users to assign tags lies in a better organisation and retrieval of multimedia items. Ames and Naaman (2007) show that Flickr users annotate their photos mainly due to social motivations. Incentives lie in a better general organisation i.e., in photo pools, for search and self-promotion, in a better organisation for their own retrieval, and in benefits for social communication such as adding context for friends and family.

In an analysis of the user tags in 100,000 Flickr images, Bischoff et al. (2010) show that Flickr users mainly annotate images with tags referring to the topic (48%), location (27%), and opinions (7%) of images. Further, they analyse query logs of AOL image search and compare them to the distribution of tags on Flickr. While many queries ask for special topics in images, they also note that there are many further subjective image queries for terms like *funny* or *sexy*, as the tag distribution in Flickr shows. Bischoff's results on Flickr tag distribution is also confirmed by Sigurbjörnsson and van Zwol (2008), who analyse a random snapshot of 52 million Flickr images with tags and categorise them into the most common WordNet categories. The distribution of tags in the location category is similar to Bischoff's results with 28%. Sigurbjörnsson and van Zwol (2008) draw a finer distinction regarding the topic category and relate 16% of the terms to artefacts or objects, 13% to people or groups, and 9% to actions or events. In their study, no affective or emotional category is defined, but 27% of the tags fall into the category *other*. Recently, Xie et al. (2010) have proposed a vocabulary of 5000 tags that are popular, visually observable, and frequently appearing on Flickr. They chose 60 tags to construct a faceted taxonomy defining possible co-occurring and mutual exclusive tags. In a user study, they verified the usefulness of these 60 tags. Results show that colour tags are not perceived to be useful, although they frequently appear on Flickr.

The tag distribution and the social incentives of tagging show that user-generated tags directly represent an information need or an expected information need of one user about the search behaviour of other users. As a result, tags can be a valuable source to derive visual concepts.

5.5. Definition of visual concepts and the Photo Tagging ontology

The definition of concepts follows the results on user studies and preferences for concepts in user-generated tags outlined above. The main categories of the ImageCLEF VCDT topics are *Content Element*, *Scene Description*, *Representation*, *Quality*, and *Emotion and Affect*. The distribution of visual topics in these categories and several subcategories is illustrated in Table 5.1. In total, 68% refers to content-based attributes, 14% is based on representational characteristics, and 12% describe emotional or affective properties. The quality terms account for 6% of the concepts and can be grouped partly into the representational attributes (technical quality) and the affective properties (subjective quality). This distribution of concepts is in line with the results of our user study, which showed 70% content-based decisions, 13% representational characteristics, and 13% affective properties in the natural grouping task. Please note that the definition of concepts is not performed completely independently of the image dataset. While the concepts are determined by the user study results and Flickr user tags, the definition process includes a visual inspection of the images, i.e., discarding concepts that are not present or rarely found in the images.

In the following, the PTO for the annotation of consumer photos is introduced. It structures the visual concepts. The visual concept definition and structuring was performed iteratively. Therefore, the PTO was extended with new visual concepts and built upon its former version over the years until it contains a total of 99 visual concepts introduced above. Three versions of the PTO exist, including the concepts illustrated in the Tables A.1, A.2, and A.3 in the appendix. The tables further denote into which categories of the ontology the concepts are placed.

Table 5.1.: Subdivision of the visual concepts in the ImageCLEF VCDT collection in categories.

Category	Subcategory	Number concepts	Percentage
Content Element	Landscape Elements	12	44.44%
	Urban Elements	3	
	General	3	
	Persons & person related	12	
	Vehicle	7	
	Animals	7	
Scene Description	Abstract	9	23.23%
	Activity	1	
	Seasons	4	
	Place	2	
	Daytime	4	
	Events	3	
Representation	General	3	14.14%
	Illumination	4	
	Art	3	
	abstract	4	
Quality	Blurring	4	6.06%
	Aesthetics	2	
Emotion and Affect	opinions	4	12.12%
	emotions	8	

5.5.1. Photo Tagging ontology: V 1.0

The PTO is an OWL ontology which is used in the ImageCLEF Photo Annotation task of 2009 as knowledge source (see Chapter 10). It consists of four top level categories and 53 visual concepts in its first version. The categories at the top level are named *Content Element*, *Scene Description*, *Representation*, and *Quality*. These categories structure the visual concepts and are understood as abstract concepts that are not directly used in annotation.

5.5.1.1. Categories and concepts

Content element: The category *Content element* covers all object-based items divided into the two subcategories *Landscape elements* and *Pictured objects*.

- Landscape elements:

This category contains several concepts from nature. This includes the concepts *Mountains*, *Water* with its specialisations *Lake*, *Sea*, and *River*, *Sky* with its sub-concept *Clouds*, and *Plants* with the sub-concepts *Flowers* and *Trees*. The concepts are structured hierarchically, so that more specific landscape elements are sub-concepts of the broader concepts. Further, they are modelled optionally, which means that all concepts can be annotated at the same time in one photo.

- Pictured objects:

The category *Pictured objects* contains the concepts `Animals`, `Food`, and `Vehicle` and the subcategory *Persons*. The concepts are modelled as optional concepts in the ontology. The concept `Animals` only describes photos of living animals, which means no soft toys or comics of animals are annotated. However, animals are regarded as representing the concept `Animals` if the whole picture is a painting.

The category *Persons* is a subcategory of *Pictured objects*. It is divided into four groups: the concept `Single` describes one person, the concept `Small Group` is used for two to five persons, the concept `Big Group` describes more than five persons, and the concept `No Person` is utilised in case the image shows no persons. Persons are defined to represent a person concept if a reasonable part of the person is visible (a hand or a belly is not representative for a person). Also, comic persons are not regarded as persons as long as the whole photo is not a comic or painting. Dolls or shadows of persons and persons far in the background of the photo are not considered as persons. The person concepts are modelled disjointly.

Scene description: The category *Scene description* covers subcategories and concepts that illustrate the place, time, or activity of the photo content. It is subdivided into the five subcategories *Abstract categories*, *Activity*, *Place*, *Seasons*, and *Time of day*.

- Abstract categories:

This category describes a rough categorisation into several scenes. It mostly contains broader concepts that summarise scene types. These concepts are modelled as optional. The concept `PartyLife` describes photos depicting a party, a concert, a celebration, or those that were taken in a club. The concept `Family and Friends` characterises photos that show elements commonly associated with family, friends, or colleagues, e.g., several persons sitting on a sofa or dining. The concept `Landscape/Nature` is used to annotate landscape images. It refers to a landscape scenery and should not be used in photos depicting close-ups of plants or animals. The concept `Buildings/Sights` illustrates images depicting buildings or sights in a rather close-up fashion. `CityLife` includes photos showing the life in cities or a skyline of a city. The concept `Desert` is used to describe photos that depict a desert. `Snow or Skiing` is annotated in photos about winter sports, winter holidays, or snowy landscapes. The concept `Beach/Holidays` is applied if photos depict the beach or express the feeling of holidays at a beach.

- Activity:

This category only contains the concept `Sports`. This concept is annotated if the photo shows people doing sports or if it depicts sport requisites.

- Place:

The concepts of the category *Place* are modelled disjointly. An image depicts a scene that takes place `Indoor` or `Outdoor`, or it is not visible from the photo where the scene takes place (`No Visual Place`).

- Seasons:

The category *Seasons* includes the sub-concepts `Spring`, `Summer`, `Autumn`, `Winter`, and `No Visual Season`. These concepts are modelled disjointly to each other. One of the four

seasons is depicted visually in the photo or no visual season can be recognized. Recognizing the seasons is in some cases rather subjective. It includes implicit knowledge by the annotators to recognize seasonal characteristics.

- Time of day:

The category *Time of day* contains the concepts *Day*, *Night*, *No Visual Time*, *Sunny*, and *Sunset/Sunrise*. The first three concepts are modelled disjointly, so a photo depicts either day or night scenes or no visual hint on the time. *Sunny* and *Sunset/Sunrise* are optional concepts. As *Sunset/Sunrise* refers to a time in the change of day and night, photos should be annotated consistently as *Day* as long as it is not nearly dark.

Representation: This category describes how image content is represented. It is composed of the concepts *Canvas*, *Macro Image*, *Still Life*, and *Portrait*, and the subcategory *Illumination*. The concepts are modelled as optional. The concept *Canvas* is used for all photos that depict no real-world scenery but graphics, illustrations, comics, or graffiti. The concept *Macro Image* is utilised for photos showing a macro or close-up view of something. The concept *Still Life* refers to the arrangement of mostly inanimate objects in an artificial way. *Portrait* photos show a portrait of persons or animals.

- Illumination:

The concepts *Neutral Illumination*, *Overexposed*, and *Underexposed* describe the illumination in photos. They are modelled disjointly, so each photo needs to have one annotation describing the illumination. It is not differentiated between accidental or intended illumination properties. Low-key photos are therefore annotated as *underexposed* and high-key photos as *overexposed*.

Quality: This category describes the quality degree of photos. It contains the two subcategories *Blurring* and *Aesthetics*. The first one describes solely technical aspects of the photo composition that can be objectively determined, while the second one includes rather subjective decisions.

- Blurring:

This category describes if photos are blurry and which blur characteristic is present. It includes the concepts *Motion Blur*, *Out of Focus*, *Partly Blurred/Depth of Focus*, and *No Blur*. These concepts are modelled disjointly.

- Aesthetics:

This category contains concepts that are rather subjective in assessment. It consists of the concepts *Overall Quality*, *Aesthetic Impression*, and *Fancy*. *Overall Quality* should be annotated if the photo is high quality or represents a very good style of photography. *Aesthetic Impression* is annotated if the photo is aesthetic or if it stands out in contrast to “normal” photos. *Fancy* is used if the photo content or the way of taking the photo was extraordinary. These concepts are modelled as optional.

5.5.1.2. Relations

The structuring of visual concepts in an ontology implies the definition of relations among them. Naturally, the ontology provides hierarchical relationships. Further, some of the concepts are modelled disjointly in the PTO. As a result, a reasoning system reports an error if two disjoint concepts are assigned to the same photo. Besides these relations, other arbitrary relations are defined. To

give an example, the object property `hasAnimals` is defined to have a domain of `MediaObject` (the super class of `photos` and `videos`) and a range of `PicturedObjects.Animals`. The `hasPerson` property belongs to the domain `MediaObject` and the range `PicturedObjects.Persons`. Then, the concept `Portrait` can be restricted so that it only can be assigned to a photo if the photo already has a relation `hasAnimals` or `hasPersons` with a further restriction on the concepts `Single`, `Small Group`, or `Big Group`. Additionally, ontologies allow to define object properties as *functional*. With this characteristic, the user is forced to annotate a photo with at least one annotation from a group of concepts. For example, the object property `hasIllumination` includes this functional characteristic. As a result, each photo needs to have a relation `hasIllumination`. As the illumination concepts are also defined to be disjoint, these two mechanisms guarantee that each photo instance is annotated with an illumination concept and that it is only annotated with exactly one of the illumination concepts at the same time.

5.5.2. Photo Tagging ontology: V 2.0

The second version of the PTO was defined in 2010 and is used in the ImageCLEF Photo Annotation task of 2010 (see Chapter 10). In total, 42 visual concepts and a few structural categories have been added. The concept `Canvas` was removed as its understanding was ambiguous. It is replaced by the concept `Visual Arts` which subsumes the concepts `Graffiti` and `Painting`. Some of the former categories are extended by new concepts. The concept `Architecture` is introduced in *Abstract Categories* and the concepts `Park/Garden` and `Rain` are included in *Landscape elements*. The category *Pictured objects* is augmented with the concepts `Toy` and `Musical Instruments`, the category *Illumination* now contains a concept `Shadow`, and the *Persons* category is enlarged by the concept `bodypart`. Further, some visual concepts are specialised by sub-concepts. The `Animals` concept is specialised into different animal species. This includes the concepts `Dog`, `Cat`, `Bird`, `Horse`, `Fish`, and `Insect`. The `Vehicles` concept is specialised into several kind of vehicles. This includes the concepts `Car`, `Bicycle`, `Ship`, `Train`, `Airplane`, and `Skateboard`. Two top level categories are added to the PTO. The category *Urban Elements* is defined as counterpart to the category *Landscape Elements*. It includes the concepts `Street`, `Church`, and `Bridge`. The concept `Fancy` was moved from *Aesthetics* to the new top level category *Impression* and complemented with the concepts `Boring` and `Cute`. Finally, several new categories are defined. The category *Events* is placed as subcategory in *Scene Description* and contains the concepts `Travel`, `Work`, and `Birthday`. The category *Impression* is defined as subcategory of *Representation* and includes the concepts `Natural`, `Abstract`, `Technical`, and `Artificial`. Further, two person specific categories are added. The category *Gender* includes the concepts `male` and `female` while the category *Age* consists of the concepts `Baby`, `Child`, `Teenager`, `Adult`, and `Old Person`.

5.5.3. Photo Tagging ontology: V 3.0

The third version of the PTO is used in the ImageCLEF Photo Annotation task 2011 as an additional knowledge resource. It includes nine sentiment concepts. Therefore, the category *Impression* is further divided into the categories *Expressed* and *Felt*, which determine the sentiment the image conveys and the sentiment the viewer senses, respectively. The category *Expressed* includes the concepts `Active`, `Euphoric`, `Happy`, `Calm`, `Inactive`, `Melancholic`, `Unpleasant`, and `Scary`. These concepts are derived from the circular affect model proposed by Russell (1980). The category *Felt* contains the sentiment `Funny`. Further, the concepts `Boring`, `Cute`, and `Fancy` are moved to this subcategory. The three concepts `No Visual Place`, `No Visual Season`, and `No Visual Time` are removed from the ontology. These concepts have a large occurrence in the dataset and were criticized in the community as they contain a negation and do not directly describe a visual occurrence of an entity or scene. See also Figure A.1 in the appendix.

5.6. Comparison to visual concepts of other test collections

The visual concepts of other test collections for image and video annotation have been introduced in Chapter 4.1. In the following, the coverage of the visual concepts of nine collections is compared to the concepts of the PTO. In particular, this includes the collections LSCOM, Caltech-256, NUS-Wide, IAPR-TC-12, VideoAnnEx, SUN-397 (i.e., the 397 well sampled categories), ImageNet, MIR Flickr, and PASCAL VOC (classification task 2011). The results are presented in Table 5.2. In cases where the collection provides an ID for the specific concept, the ID of the matching concepts is listed. Otherwise, an "X" denotes the presence of the concept in the respective collection. The brackets indicate that there is a similar concept in the collection, which is, however, not perfectly matching. Finally, the total number of equal concepts is shown (including near matching concepts) and it is illustrated how big the coverage of these concepts is in comparison to the PTO. The concepts of the PTO, which are not contained in any of the other collections, are printed in bold.

Results show the highest overlap between LSCOM and PTO with 55%. The basic difference lies in the quality concepts and the subjective concepts such as *happy* or *boring*. Naphade et al. (2006) state that they chose concepts according to the possibility for automated approaches to detect them. This feasibility was assessed in a 5 years horizon. In the last years, automated approaches show improved detection performances for scene and object annotation. The author therefore believes that it is time to provide a challenging dataset including affective concepts in order to address the user needs regarding image indexing, and in order to investigate how far automated approaches can support these requests. The quality concepts and the event concepts can also not be found in LSCOM. The ImageNet and the IAPR-TC-12 collections cover the PTO concepts with 48% and 44%, respectively. The author expects the ImageNET collection to contain more or less all concepts when it is ready. Only the concepts describing adjectives such as *overexposed*, or *natural* are not present in the noun structure of WordNet which is used as concept hierarchy in ImageNet. However, some of the adjectives might be expressed as nouns, such as *happy* with *happiness*. Nonetheless, it has to be noted that the images in ImageNet visually differ from the ones in the ImageCLEF collections. There is an overlap of 33% with the concepts of the NUS-WIDE test collection. The NUS-WIDE collection especially includes more animal species, different types of events and more landscape elements. The Caltech-256 collection contains very specific objects. To give an example, the set includes a concept *golden-gate bridge*, but no concept just asking for *bridges*. The overlap to the PTO is very low with 16%, especially if the comparison is performed strictly, this reduces to 6%. Please note that while the coverage of the original MIR Flickr concepts and the VOC concepts in comparison to PTO is low (23% and 11%), these collections do not contain many concepts. The MIR Flickr concepts are included in the PTO for 22 out of 23 concepts in contrast to VOC, for which only 11 out of 20 concepts are contained. The high inclusion rate of the original MIR Flickr concepts is natural, as these were also derived from the most frequent Flickr user tags of the set.

All in all, no collection contains the concepts of the representational category. In a few cases, only the concepts *portrait* and *still image*, as a similar concept to *still life*, occur. Further, the more subjective concepts, such as the emotional concepts, events, or the seasons, cannot be found in related work. Especially, the quality concepts are motivated by user studies, such as described in Rodden and Wood (2003). In addition, ways to sort out *boring* or *unpleasant* images are indispensable. The PTO with its holistic view on image analysis provides a strong means to sort different kind of concepts and to assess images on a broad basis.

Further, the categories of the PTO have been defined almost independently of the MIR Flickr image set on which they are applied to. Some categories are inspired by visually inspecting the photos or by the Flickr tags. Still, the dataset was not crawled by having specific categories in mind (in contrast to most other available datasets that perform a keyword search on several Web

Table 5.2.: Comparison of the visual concepts of the PTO to concepts in related datasets.

No	Concept	LSCOM ID	Caltech-256 ID	NUS-Wide	IAPR-TC-12	VideoAnnEx	SUN-397	ImageNET	MIR Flickr	VOC
0	Partylife	039						x		
1	Family_Friends	580						x		
2	Beach_Holidays	087		x	x	x	x	x		
3	Building_Sights	226		x	x	x	x	x		
4	Snow	024		x	x	x	x	x		
5	Citylife	068		x	x	x	x	x		
6	Landscape_Nature	153		x	x	x	x	x		
7	Sports	415								
8	Desert	230		x	x	x	x	x		
9	Spring									
10	Summer									
11	Autumn									
12	Winter									
13	Indoor	(398)				x	(x)		(x)	
14	Outdoor	224				x	(x)			
15	Plants	235		x	x	x	x	x		(x)
16	Flowers	307	(204)	x	x	x	x	x	x	
17	Trees	435	(154)	x	x	x	x	x	x	
18	Sky	207		x	x	x	x	x	(x)	
19	Clouds	276		x	x	x	x	x	x	
20	Water	(209)	(241)	x	x	x	(x)	x	(x)	
21	Lake	335		x	x	x	x	x	x	
22	River	078		x	x	x	x	x	(x)	
23	Sea	356		x	x	x	x	x	x	
24	Mountains	236		x	x	x	x	x		
25	Day	(290)								
26	Night	352		x					x	
27	Sunny	423								
28	Sunset_Sunrise									
29	Still_Life	(412)/(810)							(x)	
30	Macro									
31	Portrait	(150)							(x)	
32	Overexposed									
33	Underexposed									
34	Neutral_Illumination									
35	Motion_Blur									
36	Out_of_focus									
37	Partly_Blurred									
38	No_Blur									
39	Single_Person	410	(253)	x	x	x		x		x

continuing on next page ...

5. Topic definition: Which concepts do users prefer to organise photo sets?

No	Concept	LSCOM ID	Caltech-256 ID	NUS-Wide	IAPR-TC-12	VideoAnnEx	SUN-397	ImageNET	MIR Flickr	VOC
40	Small_Group	(686)	(159)		(x)	x		x	(x)	
41	Big_Group	(686)	(159)		(x)	x		x	(x)	
42	No_Persons									
43	Animals	201		x	x	x		x	(x)	
44	Food	187		x	x	x		x	(x)	
45	Vehicle	108		x	x	x		x	(x)	
46	Aesthetic_Impression									
47	Overall_Quality									
48	Fancy									
49	Architecture									
50	Street	418		x	x	x	x	x		
51	Church	532		x	x	x	x	x		
52	Bridge	260	(086)	x	x	x	x	x		
53	Park_Garden	101/598		x	x	x	x	x		
54	Rain	387								
55	Toy			x	x			x		
56	MusicalInstrument							x		
57	Shadow									
58	bodypart									
59	Travel	155/339								
60	Work									
61	Birthday									
62	Visual_Arts					x		x		
63	Graffiti									
64	Painting									
65	artificial	(320)				x				
66	natural									
67	technical									
68	abstract									
69	boring									
70	cute									
71	dog	292	056	x	x			x	x	x
72	cat	270		x	x			x	x	x
73	bird	257	113	x	x			x	x	x
74	horse	322	105	x	x			x	x	x
75	fish	305	(087)	x	x			x	x	x
76	insect									
77	car	221	(252)	x	x	x	(x)	x	x	x
78	bicycle	256	146/224			x		x	x	x
79	ship	405	(197)	x	x	x		x	x	x
80	train	386		x	x	x	x	x	x	x
81	airplane	218	251	x	x	x	(x)	x	x	x
82	skateboard	185	185	x	x	x		x	x	x

continuing on next page ...

5.6. Comparison to visual concepts of other test collections

No	Concept	LSCOM ID	Caltech-256 ID	NUS-Wide	IAPR-TC-12	VideoAnnEx	SUN-397	ImageNET	MIR Flickr	VOC
83	female	103			x			x	(x)	
84	male	104			x			x	(x)	
85	Baby	247			x			x	(x)	
86	Child	273			x			x	(x)	
87	Teenager	828			x			x		
88	Adult	181						x		
89	old_person	359						x		
90	happy									
91	funny									
92	euphoric									
93	active							(x)		
94	scary									
95	unpleasant									
96	melancholic									
97	inactive							(x)		
98	calm									
number overlap		56 55,44%	16 15,84%	33 32,67%	44 43,56%	28 27,72%	18 17,82%	48 47,52%	23 22,77%	11 10,89%

image search engines and perform a cleansing of the photos per concept later). The crawling of the MIRFlickr dataset focused on the “interestingness” and on the copyright of images. Therefore, the dataset can be assumed to contain less selection bias than other datasets and to present a variety of negative examples for concepts. A recent work of Torralba and Efros (2011) refers to the bias of test collections and points out the need to reduce selection bias, labelling bias, or negative set bias. They impressively show that most datasets can be immediately recognized just by visually inspecting photos of one concept of different datasets.

Further, the number of images per concept varies considerably in the ImageCLEF VCDT collection. Often, artificial distributions of concept instances are present in test collections due to the process of first defining the concepts and then collecting a special amount of data per concept. Usually, this amount is equal for different concepts, which neglects real-world appearance.

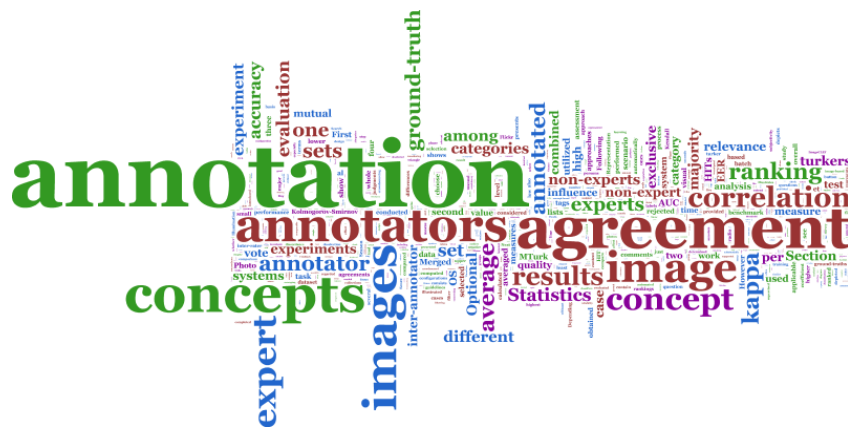
5.7. Summary

This chapter describes the process of visual concept definition and the definition of the PTO for the annotation of personal photos. Related work on how users describe images and what requests they pose on image indexing approaches is reviewed. The results of the user study on how people organise photo collections are in line with the results of previous studies, but verify the outcomes on a larger image database containing personal photos. Further, motivations for tagging on Flickr and a review on the topics of these tags has been performed. This analysis shows that user tags directly incorporate the information need of a user, or respectively, the expected information need of other users. This makes user-generated tags a valuable source for deriving visual concepts. Finally, the visual concepts of the ImageCLEF VCDT collection are defined and compared to related test collections. The visual concepts include content-based terms, representational terms, and affective terms, and therefore follow results of user studies on image descriptions. This generic view on image properties determines one of the strengths of the visual concept lexicon. Related work mainly deals with content-based attributes, while especially the affective and the quality aspects are completely missing. This stands in direct contrast to the user’s information need.

A limitation of the PTO lies in the rather subjective structuring of the concepts. The work could be improved by asking subjects to sort the visual concept lexicon in a graph-like structure, and to define relations among the concepts. An experiment on how closely the structure of the PTO relates to human semantic relatedness estimation is detailed in Chapter 8.5. Furthermore, 99 visual concepts are only a small number of concepts in comparison to other test collections as the LSCOM dataset. On the other hand, the ImageCLEF VCDT collection is used in the ImageCLEF benchmark as test collection (see Chapter 10). In comparison to the number of concepts in other evaluation campaigns, such as the TRECVID High-level feature extraction task or PASCAL VOC classification task, the PTO defines a reasonable number of concepts.

Finally, the process of collecting images and defining visual concepts is different from related work. While usually the concept lexicon exists before images are collected, in the case of the ImageCLEF VCDT dataset, this process is decoupled, and the images have been collected first. This approach is much closer to reality and poses new challenges as objects are not necessarily centred in the image and the distribution of images per concept varies considerably. Further, the visual concept lexicon is applicable to different use cases as image search or photo management. This is directly supported by concepts such as `boring` or `out of focus` which are of great interest in photo management software in order to automatically reject photos. Therefore, the visual concept lexicon and the PTO are highly valuable resources for next-generation image annotation approaches.

6. Relevance assessment: Concept subjectivity and inter-annotator agreement



This chapter focuses on the second issue with test collections, the relevance judgements. The manual assessment of relevance is a costly business in terms of time and effort, and is usually performed by experts. Optimally, more than one judgement per document is obtained to ensure high quality annotations and to minimise subjectivity. However, often no expert knowledge is needed to judge relevance. In this context, I investigate the subjectivity of concepts assessed by experts and non-experts and the influence of different relevance sets on the evaluation of image annotation approaches. This study especially focuses on the question of whether relevance assessment for image annotation can be outsourced and obtained with a crowdsourcing approach. The chapter is structured as follows. After a brief motivation, Section 6.2 explains the setup of the study on inter-annotator agreement by illustrating the dataset and the relevance acquisition process. Section 6.3 details the methodology of the experiments and introduces the relevant background. Next, Section 6.4 presents the results of the four experiments, followed by a discussion in Section 6.5. Section 6.6 introduces an alternative to manual relevance assessment which makes use of user-generated tags and a supervised machine learning approach. Finally, the chapter concludes with a brief summary in Section 6.7. The study on inter-annotator agreement is published in Nowak and R ger (2010).

6.1. Motivation

In IR and machine learning, golden standard databases play a crucial role. They are required to train supervised classification approaches, and they allow to compare the effectiveness and quality of systems on a defined test collection. Depending on the application area, creating large, semantically annotated corpora from scratch is a time- and cost-consuming activity, as especially the assessment of relevance is expensive. Usually, experts review the data and perform manual

annotations. Often, different annotators judge the same data, and the inter-annotator agreement is computed among their judgements to ensure quality and minimise subjectivity.

The difficulty of relevance assessment, including issues on consistency, completeness, and reusability, has been extensively discussed in Chapter 4.2. Related work on assessing inter-annotator agreement and on adopting crowdsourcing approaches for relevance assessment have been detailed. However, the analysis of related work reveals that no study has presented an extensive evaluation on the quality of crowdsourced annotations in comparison to expert annotations and on the implications in using a crowdsourced collection in image annotation benchmarks. Studies either investigate the effect of crowdsourced relevance judgements for textual classification approaches or obtain relevance judgements for visual collections without studying their effect on evaluation. The goal of this work is twofold. First, the amount of differences among several sets of expert annotations is investigated in order to see whether repeated annotation is necessary, and whether it influences performance ranking in a benchmark scenario. Second, the reliability of non-expert annotations is explored and it is analysed if the reliability of the annotations is sufficient to be employed as ground truth in a benchmarking campaign. Therefore, four experiments on inter-annotator agreement are conducted and applied to the annotation of an image corpus with multiple labels. Following the work of Sorokin and Forsyth (2008) and Deng et al. (2009), image annotations are obtained utilising MTurk. In the experiments, these annotations are acquired on an image-based level for a multi-label scenario and compared to expert annotations. Extending the work performed on inter-annotator agreement in Alonso and Mizzaro (2009) and Brants (2000), I do not only analyse the inter-rater agreement, but study the effect of multiple annotation sets on the ranking of systems in a benchmark scenario.

6.2. Experimental setup

The experiments are conducted on a subset of 99 images from the ImageCLEF VCDT test collection. This set comprises a total of 18,000 images and it was utilised for the Photo Annotation task in the ImageCLEF benchmark from 2009 to 2011 (see Chapter 10). All images were annotated with 53 visual concepts (Table A.1) by expert annotators of the Fraunhofer Institute for Digital Media Technology (IDMT) for the annotation challenge in 2009. In that year, 19 research teams submitted a total of 74 run configurations. The 99 images utilised in the experiments on inter-annotator agreements are part of the test set of the task. Consequently, the results of 74 run configurations in automated annotation of these images can serve as a basis for investigating the influence of different ground truth sets on system ranking.

In the following, the process of obtaining expert annotations is illustrated by outlining the design of the annotation tool and the task the experts had to perform. Subsequently, the acquisition process of obtaining ground truth from MTurk is detailed, including the design of the annotation template and the rejection process.

6.2.1. Collecting data of expert annotators

The set of 99 images was annotated 11 times by 11 expert annotators from the Fraunhofer IDMT research staff with 53 concepts. The expert annotators were provided with a definition of each concept, including example photos (see Chapter 5). The 53 concepts to be annotated per image were ordered into several categories. In principle, there were two different kinds of concepts: optional concepts and mutual exclusive concepts. For instance, the category *Place* contains three mutual exclusive concepts, namely *Indoor*, *Outdoor*, and *No Visual Place*. In contrast, several optional concepts belong to the category *Landscape Elements*. The task of the annotators was to choose exactly one concept for categories with mutual exclusive concepts and to select all

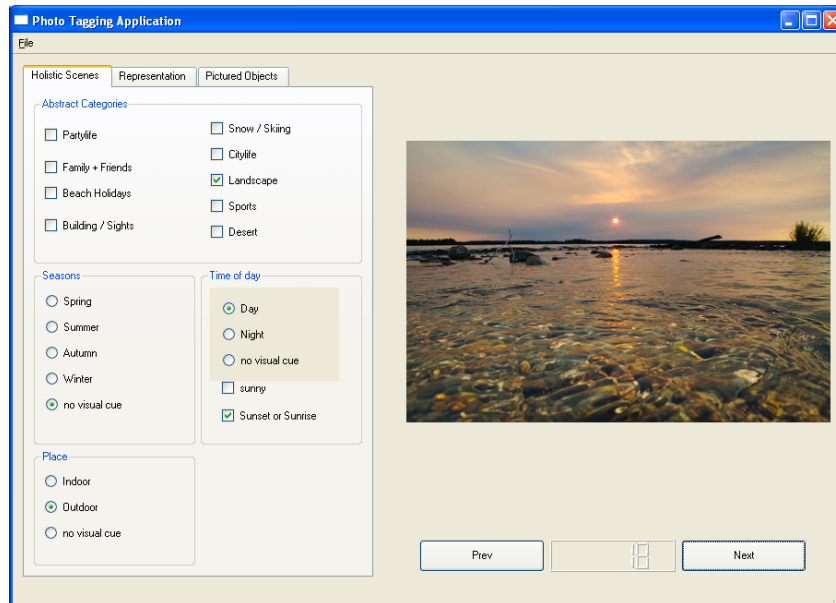


Figure 6.1.: Annotation tool that was used for the acquisition of expert annotations.

applicable concepts for optionally designed concepts. All photos were annotated at an image-based level. The annotators assessed an image with all applicable concepts and then continued with the next image.

Figure 6.1 shows the annotation tool that was delivered to the annotators. The categories are ordered into the three tabs *Holistic Scenes*, *Representation*, and *Pictured Objects*. All optional concepts are represented as check boxes and the mutual exclusive concepts are modelled as radio button groups. The tool verifies, if for each category containing mutual exclusive concepts, exactly one was selected before storing the annotations and presenting the next image.

6.2.2. Collecting data of non-expert annotators

The same set of images that was assessed by the expert annotators was distributed over MTurk and annotated by non-experts in the form of mini-jobs. The design of these HITs is similar to the expert annotation tool. Each HIT consists of the annotation of one image with all applicable 53 concepts. It is arranged as a survey and structured into three sections. The section *Scene description* and the section *Representation* each contains four questions, the section *Pictured objects* consists of three questions. On the top of each section, the image to be annotated is presented. The repetition of the image ensures that the turker can see it while answering the questions without scrolling to the top of the document.

Figure 6.2 illustrates the survey for the section *Representation*. Each HIT starts with a set of instructions followed by the actual task. As a consequence, the guidelines should be very short and easy to understand. In the annotation experiment, the following annotation guidelines were posted to the turkers. These annotation guidelines are far shorter than the guidelines for the expert annotators and do not contain example images.

- Selected concepts should be representative for the content or representation of the image.
- Radio button concepts exclude each other. Please annotate with exactly one radio button concept per question.



Representation

1. How is the image illuminated? (Choose the most applicable)
 - Overexposed
 - Underexposed
 - Neutral Illumination
2. Is the image blurred? (Choose the most applicable)
 - Motion Blur
 - Out of focus
 - Partly Blurred / Depth of focus
 - No Blur
3. How is the content of the image represented? (choose all applicable)
 - Portrait
 - Macro Image
 - Still Life
 - Canvas
4. Is the image ...? (choose all applicable)
 - of a high grade of overall quality?
 - aesthetic?
 - fancy?

Figure 6.2.: Section “Representation” of the survey.

- Check box concepts represent optional concepts. Please choose all applicable concepts for the image.
- Please make sure that the information is visually depicted in the images (no meta-knowledge).

The experiment at MTurk was conducted in two phases. In the first phase (also considered as test or validation phase), each of the 99 HITs was assigned five times, which resulted in 495 annotation sets (five annotation sets per image). One HIT was rewarded with 5 Cents. The second phase consisted of the same 99 HITs that were annotated four times by the turkers. Thus, 891 annotation sets were obtained altogether. The work of all turkers that did not follow the annotation rules was rejected. As the review of the annotation correctness is difficult and subjective, the rejection process was conducted on a syntactical level. Basically, all images in which at least one radio button group was not annotated were rejected, as this clearly violates the annotation guidelines. Overall, 94 annotation sets were rejected that belong to 54 different images (see Figure 6.3 on top). As a maximum, for one image five HITs and for two others four HITs were rejected. Looking at these images in Figure 6.3, no obvious reason can be found why so many turkers did not answer all categories of these images. Figure 6.3 illustrates the amount of images that were annotated per turker at the bottom. In both batches, there was one turker who almost annotated the whole set of 99 images.

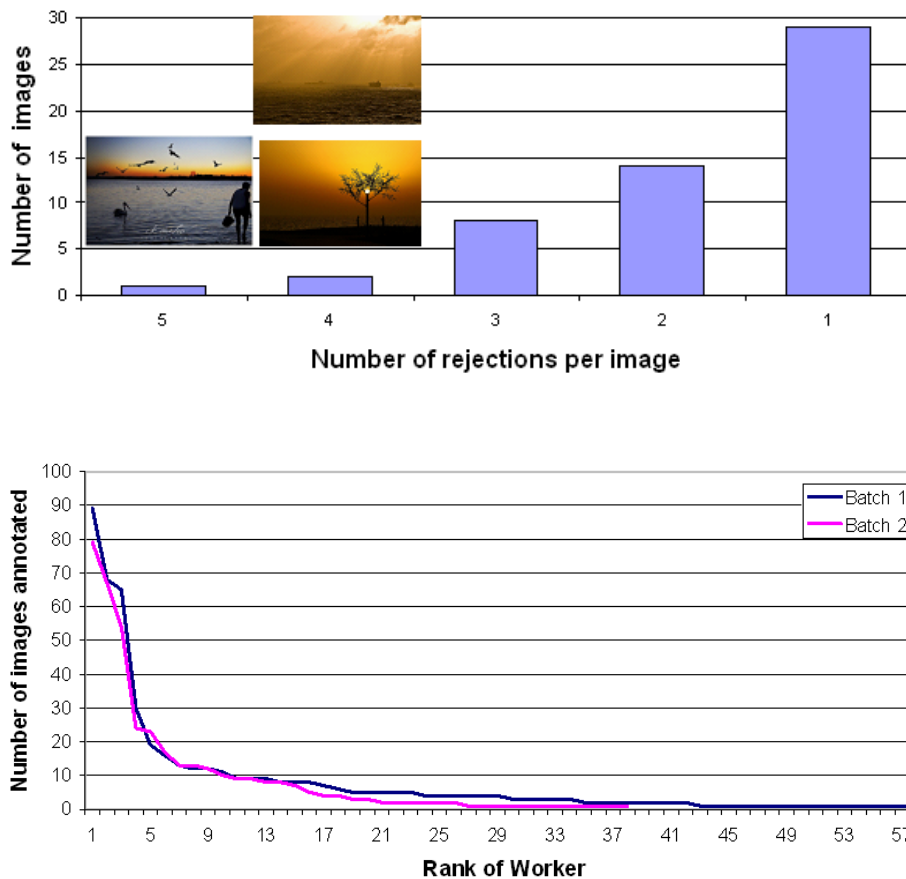


Figure 6.3.: At the top, the number of rejections per image is plotted for all rejected images. At the bottom, the amount of images annotated by each turker is illustrated.

Statistics of the first batch: The first batch, consisting of 495 HITs, was completed in about 6 hours and 24 minutes. In total, 58 turkers worked on the batch and, on average spent 156.4 seconds per HIT. 58 annotation sets were rejected, which corresponds to 11.72% of the batch. The time spent to annotate these images was on average 144.4 seconds, which does not substantially differ from the overall average working time. In all rejected images, at least one category with mutual exclusive concepts was not annotated. The first round also served as validation phase. Results were analysed to check whether there was a misconception in the task. In the survey, the category *Time of Day* consists of mutually exclusive and optional concepts at the same time. The turker was supposed to choose one answer out of the radio buttons *Day*, *Night* and *No visual time* and optionally could select the concepts *Sunny* and *Sunset or Sunrise*. For this category, it did not seem clear to everybody that one radio button concept had to be selected. As a result, the description for this category was rendered more precisely for the second round. The rest of the survey remained unchanged.

Statistics of the second batch: The second batch was published four days after the first. Its 396 HITs were completed in 2 hours and 32 minutes by 38 workers. 9.09% of the HITs had to be rejected, which equals 36 HITs. On average, 137.5 seconds were needed for the annotation of one image. The rejected images were annotated in 117.8 seconds on average. Six workers worked on both batches. MTurk arranges that several assignments per HIT are not finished by

Table 6.1.: Categorisation of comments posted by the turkers.

Content of photo	About work	Feelings about photo	Quality of photo	Other
Cupcakes	Dolls aren't persons right	Cute And nice	Color effect can be better	ha haa Sure
Looks Like a dream land	really nice to work on this. this is very different and easy. Answer for Picture Objects 2 and 3 are not fitting correctly.	Just Beautiful Thats Creative I really like this one has nice composition.	Interesting good representation for Logo or....	For what purpose is this useful?

the same turkers. However, as the second batch was published as a new task some days later, it was possible that the same turkers of the first round also worked on the second batch. All in all, there were 13 images that were annotated twice by the same person.

Feedback: Each HIT was provided with a comment field for feedback. The feedback was optional. The comments received can be classified into 1) comments about work, 2) comments about the content of the photo, 3) comments about the quality of the photo, 4) comments about feelings concerning a photo, and 5) other comments. In Table 6.1, an excerpt of the comments is listed. Notably, no negative comment was posted and the turkers seemed to enjoy their task.

6.3. Evaluation design

In total, four experiments are conducted to assess the influence of expert annotators on the system ranking and to examine whether the annotation quality of non-expert annotators is good enough to be utilised in benchmarking campaigns:

1. Analysis of the agreement among experts

The first experiment analyses the inter-annotator agreement among the expert annotators. There are different possibilities to assess the inter-rater agreement when each annotator annotated a whole set of images, which was the case for the experts. One way is to calculate the accuracy between two sets of annotations, i.e., for all images and all concepts at once. Another way is to compare the average annotation agreement on a basis of the majority vote for each concept or for each image.

In the annotation of the images, two principal types of concepts were used, optional and mutually exclusive ones. As a result, there are two possibilities on how to count a decision:

- A decision is only performed by explicitly selecting a concept.
- A judgement is performed through the selection and deselection of concepts.

In the case of the optional concepts, it is not assured whether a deselected concept was chosen to be deselected, or whether it was only forgotten during the annotation process. In the case of the mutual exclusive concepts, the selection of one concept automatically leads to a deselection of the other ones in the same group. In this case, both the selection and deselection of a group of concepts is performed intentionally. The majority analysis of

the agreement on images and concepts takes these two paradigms into consideration and compares its results.

2. Influence of different sets of expert annotations on system ranking performance

In the second experiment, the influence of annotator sets on the performance of systems is determined. The goal is to examine to what extent different ground truths affect the ranks of systems in a benchmark scenario. Each set of expert annotations is regarded as ground truth for the evaluation of the ImageCLEF submissions. The 74 run configurations of the Photo Annotation task were trimmed to contain only the annotations for the 99 images. The evaluation considers each trimmed run and each ground truth, and calculates performance using three performance measures (Ontology Score (OS), EER, and AUC). These measures constitute the official measures of the Photo Annotation task in 2009. Moreover, the OS will be introduced in Chapter 7. In a second step, the resulting ranked lists of systems for all annotator ground truths are compared to each other with the Kendall τ correlation coefficient and the Kolmogorov-Smirnov statistics introduced in Chapter 4.6.1.

3. Analysis of the agreement between experts and non-experts

The third experiment analyses the agreement between expert and non-expert annotators. Its goal is to assess whether there is a comparable agreement between non-experts and the inter-rater agreement of experts. In general, the annotations obtained from MTurk are organised as HITs. A HIT should cover a small piece of work that is paid with a small reward. As a consequence, the major differences between the expert annotation sets and the ones from MTurk are that, at MTurk, each set of 99 images is annotated by several persons. This allows the comparison of the agreement on the labels at a concept- and image-based level, but not comparing the correlation among annotators over the whole set. The analysis of non-expert agreements considers only approved annotation sets from MTurk and uses the annotation sets from both batches combined.

In this experiment, the annotation agreement for each image is calculated in terms of example-based accuracy and compared between both groups of annotators. Further, the expert annotation set determined with the majority vote is compared to the combined annotation set of the non-experts. As in the first agreement experiment, the accuracy serves as the evaluation measure. To evaluate the inter-annotator agreement on a concept basis, the kappa statistics are utilised (see Section 6.3.3). They take all votes from annotators into account and derive an agreement value which excludes the agreement by chance. This experiment is performed for both groups of annotators, and the results for the mutually exclusive categories and the optional concepts are compared.

4. Influence of averaged expert and non-expert annotations on system ranking

Finally, the influence of the non-expert annotations on the system ranking is investigated. The combined annotations determined by the majority vote of the non-experts are used as basis for evaluation. The correlation in ranking between this ground truth and the combined ground truth of the expert annotators is computed for all three evaluation measures OS, EER, and AUC. The Kendall τ correlation coefficient and the Kolmogorov-Smirnov statistics are calculated between both rankings.

In the following, the background needed to understand the experiments is briefly explained. First, the accuracy measure to compute the inter-annotator agreement is introduced. Second, the rank correlation experiments including the performance measures for image annotation evaluation are detailed, and finally, the kappa statistics are explained.

6.3.1. Accuracy for agreement assessment

The experiments consider different measures to determine inter-annotator agreement, The accuracy between two annotated image sets U and V is defined as in Brants (2000):

$$\text{accuracy}(U, V) = \frac{\# \text{ identically tagged labels}}{\# \text{ labels in the corpus}}, \quad (6.1)$$

where *labels* refer to the annotated instance of a concept over the whole image set. The accuracy between two sets of annotations per image X_i can be calculated as follows:.

$$\text{accuracy}_{ex}(X) = \frac{1}{N} \sum_{i=1}^N \frac{\# \text{ identically tagged labels in } X_i}{\# \text{ labels in } X_i}. \quad (6.2)$$

This way of calculating the accuracy does not presume the existence of two persons that annotated the whole set of images, but evaluates the accuracy of annotations on an image-based level. It is further denoted as example-based Accuracy (accuracy_{ex}).

6.3.2. Rank correlation as a measure of inter-rater agreement

In these experiments, rank correlation statistics are used as an implicit measure to determine the agreement among different annotators. The predictions of image annotation systems are evaluated with a performance measure, and the systems are ranked according to their score for different sets of the ground truth. A high correlation in compared result lists denotes an equal ranking and points to a close annotation behaviour. The experiments consider three performance measures, the OS, EER, and AUC. The OS assesses the annotation quality on an image basis and assigns fine-grained costs for misclassifications. The score is based on structure information, relationships from the ontology, and the agreement between annotators for a concept. In contrast, the measures EER and AUC assess the annotation performance of the system per concept. In case a concept was not annotated at least once in the ground truth of the annotator, it is not considered in the final evaluation score for EER or AUC. Following the arguments outlined in Chapter 4.6.1, two tests, the Kendall τ and Kolmogorov Smirnov test, are utilised in order to assess the degree of variance of different ground truths in ranking systems.

6.3.3. Kappa statistics

Kappa statistics can be utilised to analyse the reliability of the agreement among annotators. This statistical measure was originally proposed by Cohen (1960) to compare the agreement between two annotators when they classify assignments into mutually exclusive categories. It calculates the degree of agreement while excluding the probability of consistency that is expected by chance. The coefficient ranges between 0 when the agreement is not better than chance and 1 when there is a perfect agreement. In case of systematic disagreement the kappa score can also become negative. As a rule of thumb, a kappa value above 0.6 represents an adequate annotator agreement, while a value above 0.8 is considered as almost perfect (Landis and Koch (1977)). The kappa statistic used in the following analysis is called free-marginal kappa (see Brennan and Prediger (1981) and Randolph (2005)), and can be utilised when the annotators are not forced to assign a certain number of documents to each concept. It is suitable for any number of annotators.

6.4. Results

This section presents the results of the four experiments: the first analyses the agreement among different sets of expert annotations, the second investigates the influence of the annotation sets

Table 6.2.: The confusion matrix depicts the accuracy among annotators averaged over the set of 99 images and 53 concepts. The column *Merged* contains the accuracy of single annotators to the majority votes of all annotators.

	A 2	A 3	A 4	A 5	A 6	A 7	A 8	A 9	A 10	A 11	Merged
A 1	.900	.877	.905	.892	.914	.912	.890	.894	.900	.916	.929
A 2		.885	.913	.905	.916	.915	.903	.911	.909	.925	.939
A 3			.900	.873	.902	.884	.878	.886	.892	.904	.918
A 4				.897	.926	.918	.899	.914	.915	.932	.947
A 5					.900	.917	.902	.901	.900	.911	.928
A 6						.925	.902	.918	.925	.932	.952
A 7							.900	.916	.918	.929	.945
A 8								.887	.892	.909	.920
A 9									.918	.918	.941
A 10										.919	.941
A 11											.958

on system ranking, the third compares inter-rater agreement between experts and non-experts, while the last experiment considers the influence of non-expert annotations on ranking.

6.4.1. Experiment 1: Agreement analysis among experts

For each of the 11 expert annotators, the accuracy is calculated as introduced in Equation 6.3.1 in comparison to the annotations of all other annotators. Additionally, the majority vote of all annotators is utilised as 12th ground truth, further denoted as *merged annotations*. Table 6.2 presents the results in a confusion matrix. In general, the accuracy between the annotations is very high. The overall accuracy is 0.912, with a minimum of 0.873 between annotator 3 and 5 and a maximum of 0.958 between annotator 11 and the merged annotations.

The top of Figure 6.4 presents the agreement for each concept among the eleven annotators. The majority vote determines whether a concept has to be annotated for a specific image. The percentage of the annotators that chose this concept is depicted in the figure averaged over all images. Note that by calculating agreements based on the majority vote, the agreement on a concept cannot be worse than 50%. The upper line represents the agreements on a concept averaged over the set of images when the selection and deselection of concepts is regarded as intentional. This means that if no annotator chose concept *C* to be annotated in an image *X*, the agreement is regarded as 100%. The lower line represents the case when only selected concepts are taken into account. All images in which a concept *C* was not annotated by at least one annotator are not considered in the averaging process. When only a small number of annotators select one concept, the majority vote determines true label and the agreement. This means if, e.g., nine out of 11 annotators decided not to select concept *C* in an image *X*, the agreement on this concept would be about 82%. For the concept *Snow* the lower line represents an agreement of 0%. There was no annotator that annotated this concept in one of the 99 images.

At the bottom of Figure 6.4, the agreement among annotators is illustrated averaged for each image. Again, the average per image was both calculated based on selected concepts and based on all concepts. The upper line represents the average agreement among annotators for each image when taking into account the selected and deselected concepts. The lower line illustrates

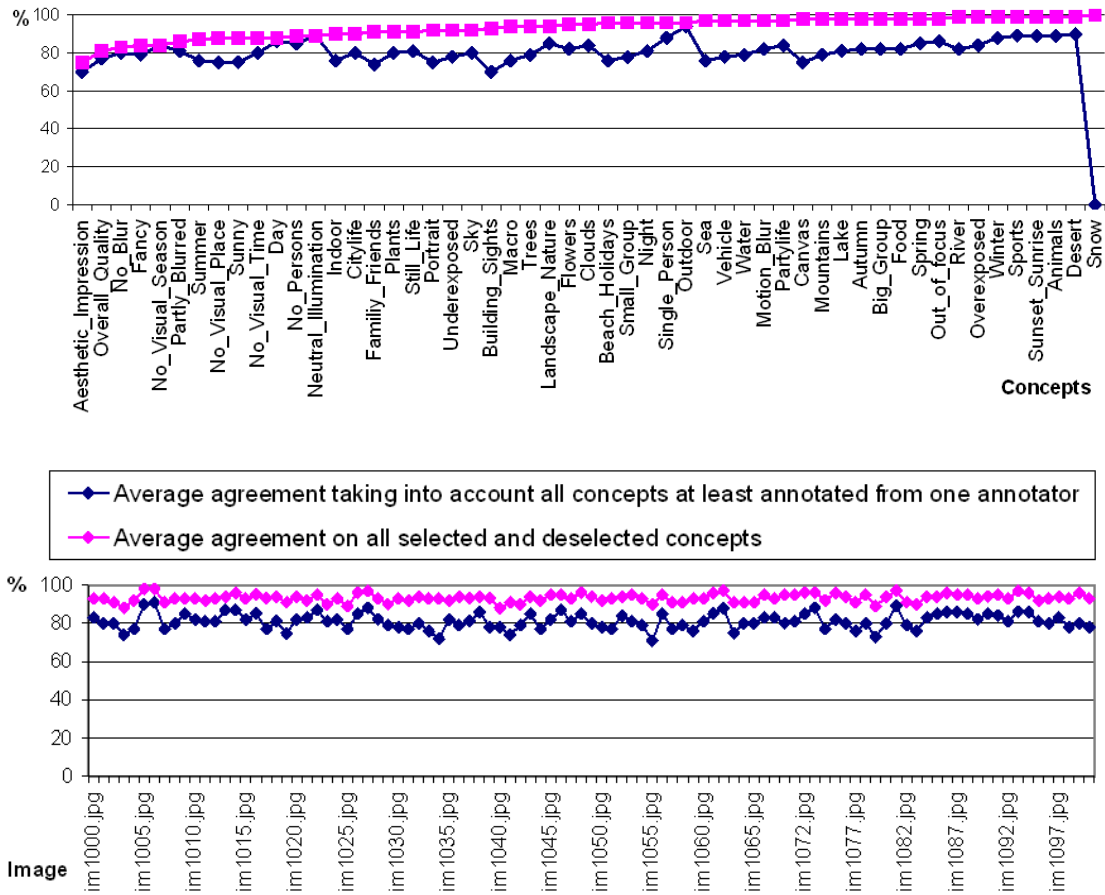


Figure 6.4.: The upper figure depicts the agreement among annotators for each concept determined by the majority vote. The lower diagram shows the inter-annotator agreement for each image.

the agreement per image when only considering the selected concepts.

All in all, the agreement among annotators in case of averaging based on the majority vote shows a mean agreement of 79,6% and 93,3% per concept, and 81,2% and 93,3% per image for selected and all concepts, respectively.

6.4.2. Experiment 2: System ranking with expert annotations

In the ranking experiment, the effect of the annotator selection on classification accuracy is investigated. Results are presented in Table 6.3, which displays the Kendall τ correlation coefficients and the decisions of the Kolmogorov-Smirnov test. The upper triangle depicts the correlations between the ranked result lists by taking into account the different ground truths of the annotators for the evaluation with the OS measure. On average, there is a correlation of 0.916 between all result lists. The list computed with the merged ground truth has an overall correlation of 0.927 with the other lists. Despite three cases, the Kolmogorov-Smirnov test supported the decision of concordance in the rankings. Overall, annotator 11 has the highest correlation to all other annotators with 0.9342, and annotator 10 has the lowest average correlation of 0.8541.

The lower triangle of Table 6.3 contains the correlation of ranks in the result lists for the evaluation measure EER. All pairs of ranked lists have a correlation of 0.883 on average. The ground truth of the annotators correlates on average to the merged ground truth with 0.906.

Table 6.3.: The upper triangle presents the Kendall τ correlation for the evaluation with the OS measure and varying ground truth, while the lower triangle shows the correlation for the evaluation with the EER score. The cells coloured in grey represent the combinations for which the Kolmogorov-Smirnov test decided on discordance.

	A 1	A 2	A 3	A 4	A 5	A 6	A 7	A 8	A 9	A 10	A 11	Merged
A 1		.938	.938	.914	.938	.884	.959	.952	.890	.816	.927	.890
A 2	.842		.964	.960	.914	.939	.951	.898	.934	.869	.960	.947
A 3	.804	.890		.969	.901	.942	.945	.897	.937	.872	.976	.948
A 4	.874	.898	.872		.892	.967	.939	.878	.959	.892	.976	.973
A 5	.872	.892	.864	.892		.859	.944	.950	.869	.792	.898	.866
A 6	.845	.927	.904	.908	.905		.910	.846	.955	.918	.951	.978
A 7	.863	.880	.872	.893	.895	.906		.928	.910	.841	.948	.917
A 8	.881	.884	.859	.896	.888	.889	.876		.850	.777	.888	.851
A 9	.819	.868	.882	.888	.875	.880	.849	.861		.892	.946	.958
A 10	.847	.867	.861	.919	.874	.881	.869	.851	.893		.872	.914
A 11	.838	.915	.894	.890	.883	.919	.898	.861	.882	.891		.954
Merged	.865	.925	.889	.919	.914	.946	.898	.905	.903	.890	.912	

For the rankings with EER, the Kolmogorov-Smirnov statistics assigned discordance in rankings in six cases. The annotator with the lowest average correlation is annotator 1 with 0.8485, and the annotations with the highest correlation on average are the ones by annotator 6 with 0.8964 (when not considering the merged annotations as having the highest correlation overall).

The results for the evaluation measure AUC are similar to the ones of the OS. The correlation between all runs is on average 0.936. The correlation between the ground truth of each annotator and the merged ground truth is on average 0.947. The lowest average correlation to all other annotations are by annotator 10 with 0.9154. The highest average correlation could be achieved by annotator 11 with 0.9452. In all cases, the Kolmogorov-Smirnov statistics supported the Kendall's τ test results for concordance.

To summarise, the results of two tests showed a high correlation of the ranked lists calculated against the ground truths of the different expert annotators. Just in a few case the test results were contradicting. Depending on the evaluation measure with which the ranked lists were computed, the average correlation varies from 0.916 (OS), 0.883 (EER) to 0.936 (AUC). One can conclude from these results that the annotators have a high level of agreement and that it does not affect the ranking of the systems substantially which annotator is chosen to provide the ground truth. For the measures OS and AUC, the same two annotators, annotator 11 and annotator 10, show the highest and the lowest average correlation to the other annotations, respectively.

6.4.3. Experiment 3: Agreement analysis between experts and non-experts

In the following, the results of the agreement analysis of non-expert annotations are presented and compared to the inter-annotator agreement of the experts. Results are obtained with the accuracy measure following Equation 6.2 and the kappa statistics.

The accuracy is computed for each image X among all annotators of that image. The averaged accuracy for each image annotated by the expert annotators is 0.81. The average accuracy among all turkers for each image is 0.79. A merged ground truth file is composed from the HITs by

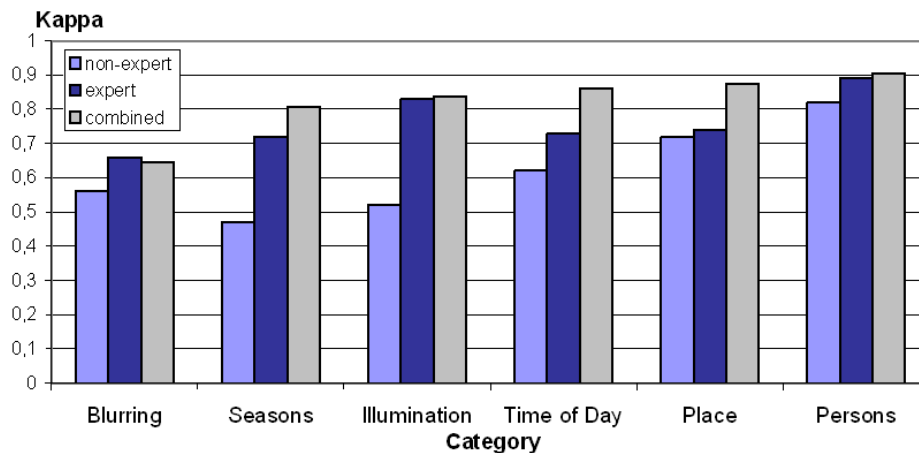


Figure 6.5.: The diagram depicts the kappa values for the mutually exclusive categories.

computing the majority vote for each concept in each image. The accuracy between the ground truth of the non-experts and the merged ground truth of the expert annotators is 0.92. In both cases, the accuracy between experts and non-experts is very high. Remembering Table 6.2, the results between the expert annotators and the merged expert annotator results are on average 0.94. Therefore, the annotations from MTurk show an accuracy that is nearly as high as that of the expert annotators.

The kappa statistics are calculated in three different configurations using the kappa calculator provided by Randolph (2008). The first configuration (denoted as *non-experts*) takes all votes of the turkers into consideration. The second, called *experts*, uses all annotations from the experts, while the third (*combined*) computes the kappa statistics between the averaged expert annotation set and the averaged non-expert annotation set. In the following, the results are presented for the categories with mutually exclusive concepts and the optional concepts.

Kappa statistics on mutually exclusive concepts: The images are annotated with six categories that contain mutually exclusive concepts. Figure 6.5 presents the results for the kappa analysis for the mutual-exclusive categories *Blurring*, *Season*, *Time of Day*, *Place*, *Illumination*, and *Persons*. The results show that the kappa value is higher for the expert annotators in all categories. The expert annotators could achieve at least a kappa value of 0.6 in each category. For the categories *Illumination* and *Persons*, an agreement higher than 0.8 could be obtained. Considering the non-expert annotations, the kappa value is only above the threshold for half of the categories. The kappa value for the categories *Season* and *Illumination* is indeed quite low. A possible reason for the category *Season* lies in the fact that the images should only be annotated with concepts that are visible in the image. In most cases, the season is not directly visible in an image and the expert annotators were trained to assign the concept *No visual season* in this case. However, the turkers may have taken guesses, which leads to a lower agreement. The kappa statistics for the combined configuration shows that the majority of the experts have a good agreement to the majority of non-experts. Except for the category *Blurring*, all agreements are higher than 0.8.

Kappa statistics on optional concepts: In the dataset, 31 optional concepts are annotated. For each optional concept the kappa statistics are exploited separately in a binary scenario for the three configurations described. Figure 6.6 presents the kappa statistics for the optional concepts. On average, the non-expert annotators agree with a value of 0.68, the experts with a value of 0.83, and the combined kappa value is 0.84. For a few concepts (*Aesthetic*, *Still Life*, *Quality*, *Macro...*)

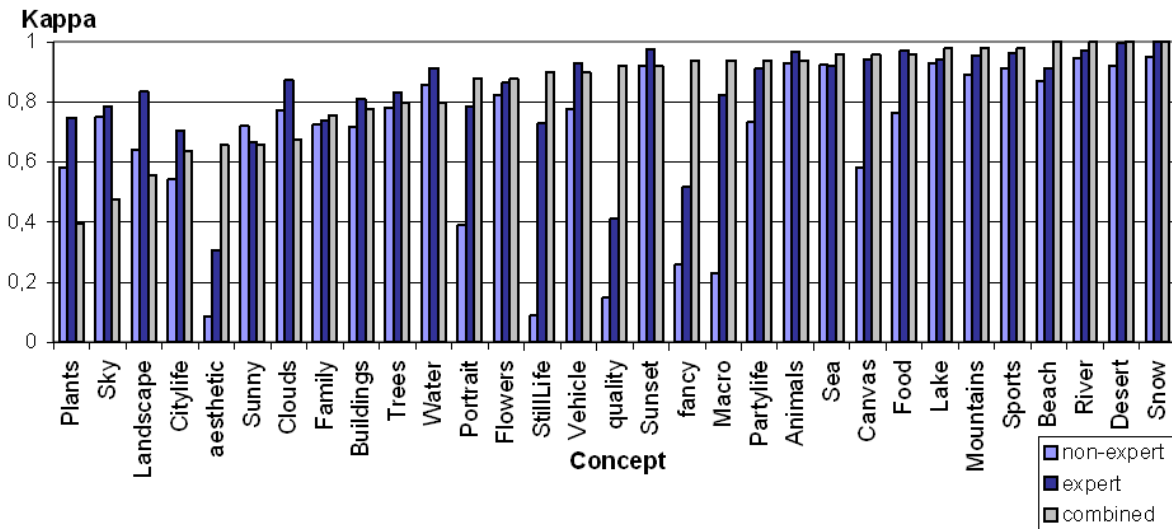


Figure 6.6.: Kappa scores for the optional concepts.

the non-expert agreement is very low. However, the combined agreement for these concepts is quite high, also slightly better on average than the agreement among experts. The results indicate that the majority vote is able to filter random answers from the non-expert annotations of most concepts and raise the averaged annotations to the level of the expert annotations. For a few concepts such as *Plants*, *Sky* and *Landscape*, the agreement among the combined annotation sets is low. These concepts are depicted quite frequently in the images, so apparently there is no major agreement about how to annotate these concepts. For other concepts, the agreement in the combined configuration is very high. Taking into account the contents of the images, the reasons for this are twofold. On the one hand, there are some concepts that are simply not often depicted in the 99 images (e.g., the concept *Snow* is not depicted at all). Thus it is correct that all annotators agree that the concept is not visible. However, the results for these images cannot be considered as a general annotation agreement for this concept, as this may change on images that depict these concepts. On the other hand, there is the problem of which annotation paradigm to apply, as illustrated in Section 6.3. If only those annotations for images in which at least one annotator selected the concept are considered, the agreement would decrease.

In contrast to the agreement results in terms of accuracy, the inter-annotator agreement evaluated with kappa statistics shows major differences between experts and turkers. While the experts were able to achieve a satisfiable agreement on concepts for most categories and optional concepts, the results of the turkers are not good in comparison. They only cross the 0.6 threshold for half of the categories. In the case of the optional concepts, 22 of 31 concepts have a non-expert agreement higher than 0.6. For other concepts, there exist major differences in agreement.

6.4.4. Experiment 4: Ranking with non-expert annotations

The last experiment explores how the different ground truths of expert annotators and turkers affect the ranking of systems in a benchmark scenario. For this experiment, the combined annotation sets obtained by the majority vote are utilised as ground truth for the system evaluation.

The Kendall τ test assigns a high correlation in ranking between the combined ground truth of the turkers and the combined ground truth of the experts. Evaluated with the OS, the correlation is 0.81, evaluated with EER, the correlation is 0.92, while utilising the AUC measure, results in a correlation coefficient of 0.93. This approximately corresponds to the correlation coefficient

the single expert annotators had in comparison to the combined expert list, as illustrated in Section 6.4.2. In this experiment, the correlation coefficient for the OS is 0.93, for the EER 0.91, and for the AUC 0.95, on average. Consequently, the ranking of the systems is affected most strongly in the case of the OS. For the EER, the non-experts annotations show even a higher correlation with the merged list than the single experts annotations. These results are supported by the Kolmogorov-Smirnov test. This test decides for concordance in case of EER and AUC, but for discordance in case of OS.

As a consequence, the results confirm that the majority vote filters random noise out of the annotations of the non-experts. However, as shown in Section 6.4.3, the agreement is still low for some concepts. It is surprising that the concept-based measures EER and AUC show such a high average correlation in the ranking of systems, even when there are a few concepts for which the annotations differ substantially among experts and non-experts. These results pose the question of whether the evaluation measures used are sensitive enough.

6.5. Discussion

To summarise, this study illustrates different experiments on inter-annotator agreement in assessing ground truth of multi-labelled images. The annotations of 11 expert annotators were evaluated on a concept- and an image-based level and utilised as ground truths in a rank correlation experiment. All expert annotators show a high consistency in annotation with more than 90% agreement in most cases, depending on the measure utilised. The kappa statistics show an agreement of 0.76 on average for the exclusive categories, and a kappa of 0.83 for the optional concepts. A further experiment analyses the influence of judgements of different annotators on the ranking of annotation systems. This indirect agreement measure exploits the fact that systems are ranked equally for the same ground truth, so that a low ranking correlation points to a low agreement in annotation. The results show that a high correlation in ranking is assigned among the expert annotators. All in all, the ranking of systems in a benchmark scenario is, in most cases, not heavily influenced by evaluating against different expert annotations. Depending on the evaluation measure used to perform the ranking, in a few combinations, a discordance between rankings could be detected. This leads to the conclusion that repeated expert annotation of the whole dataset is not necessary, as long as the annotation rules are clearly defined. However, the author suggests that the inter-rater agreement is validated on a small set to ensure quality, because depending on the concepts, the expert annotation agreement also varies.

The same experiment was conducted with non-expert annotators at MTurk. Altogether, nine annotation sets were gathered from turkers for each image. The inter-annotator agreement was not judged consistently by the different approaches. The accuracy shows a high agreement of 0.92, which is very close to the agreement among expert annotators. However, the kappa statistics report an average of 0.62 for the exclusive categories and an average of 0.68 for the optional concepts for the non-experts. The value of the exclusive concepts is close to the lower threshold of what is regarded as acceptable for annotator agreements. When comparing the averaged ground truth file of the non-experts with that of the experts, the inter-annotator agreement in terms of kappa rises to 0.84 on average. The majority vote used for generating this ground truth file seems to filter some noise out of the annotations of the non-experts. Finally, the ranking experiment shows a high correlation with combined ground truth of the non-expert annotators in comparison to that of the expert annotators. These results indicate that the differences in annotations found by the kappa statistics, even with the combined ground truth, do not significantly influence the ranking of different systems in a benchmark scenario. This poses new questions concerning the sensitivity of image annotation evaluation measures, especially in the case of concept-based evaluation measures.

6.6. Relevance from weakly labelled corpora

Despite the manual assessment of relevance by experts or non-experts in crowdsourcing approaches or gaming scenarios, there are other possibilities to automatically decide on relevant media items for a concept. Photo communities such as Flickr allow the download of photos in combination with user-generated tags. Such tags have to be revised, since they do not necessarily describe the content of the image. This problem is intensified as upload services commonly provide the possibility of uploading a set of photos and assigning the same tags to the whole set of images. Nonetheless, user-generated tags are a valuable source for test collections, or training sets for supervised classifiers. Therefore, the key challenge lies in the automated detection of images that do not match the provided tag and a consequent removal of these images from the training sets.

Some research groups focus on developing robust learning algorithms and classifiers that can deal with uncertain data, as in Ulges et al. (2008). Many methods retrieve images from the Web by textual keywords, such as in Fergus et al. (2005), Berg and Forsyth (2006), and Schroff et al. (2007). Fergus et al. (2005) utilise a PLSA approach to automatically learn topic models for retrieving object categories from Google's Image Search results. Berg and Forsyth (2006) collect images of animals via Google Text search on the category name. With the help of surrounding text, positive and negative image examples for each category are automatically selected and manually refined in a relevance feedback loop. Schroff et al. (2007) investigate different approaches to download images from the Web, automatically sort the images into drawings and natural images of the desired category, and experiment with text-based and visual-based approaches for re-ranking.

Others present approaches that automatically eliminate incorrectly labelled images as a pre-processing step and then train their classifiers on the reduced image sets. To compute a representation of a landmark site, Li et al. (2008) collect thousands of images by tags from Flickr. Outlier images are automatically sorted out through a clustering step of gist features, and through a geometric verification step by searching for a common 3D structure in the single clusters. Wnuk and Soatto (2008) retrieve target images from the Web with a k -nearest neighbour search on visual clusters, combined with a feature selection method which rejects features based on strangeness.

In Lukashevich et al. (2009), we present an automated approach to eliminate images with incorrect tags. It assumes that a reasonable proportion of images that are given the same tag relate to the desired category and share similar visual properties to a certain extent. The visual similarity is expressed through extracted colour, edge, and texture features, and used to separate targets from outliers without explicitly telling the algorithm what the negative and positive examples are. In this work, we conduct four experiments in which a one-class SVM is used to separate relevant images from irrelevant ones for a special concept. The experiments show that for all tests, a significant tendency to remove the outliers and retain the target images is present. This offers a great possibility to gather large datasets from the Web without the need for a manual review of the images. This method is especially suited for training data, as a small percentage of incorrectly labelled images does not harm the classifier to a great extent, while the demands for test collections in benchmarks are higher. An unresolved issue remains concerning the polysemy of tags. The assumption that an intrinsic visual similarity is inherent in correctly labelled images does not hold while dealing with ambiguous tags (e.g., the tag `apple` which refers to the fruit and the computer).

6.7. Summary

This chapter discusses how relevance for image annotation can be assessed while reducing the effort in terms of money and time. The study focuses on the influence of different annotation sets from experts and non-experts on the ranking of systems in a benchmark scenario. Further,

a general analysis of inter-rater agreement is conducted. Results show that while the agreement between experts and non-experts varies depending on the measure used, its influence on the ranked lists of the systems is rather small. To sum up, the majority vote applied to generate one annotation set out of several opinions is able to filter noise in judgements of non-experts to some extent. The resulting annotation set is of comparable quality to the annotations of experts. Weakly annotated corpora like the Flickr corpus allow for an automated construction of datasets. A supervised learning approach based on a one-class SVM is able to filter significant proportions of outliers and to retain the images that fit to a particular concept. This approach is particularly beneficial for gathering large training sets.

To conclude, data annotation utilising a crowdsourcing approach is very fast and cheap and therefore offers a prospective opportunity for large-scale data annotation. The results obtained in these experiments are quite promising when repeated labelling is performed and support results of other studies on distributed data annotation (Sorokin and Forsyth (2008), Alonso and Mizzaro (2009), and Snow et al. (2008)). However, as the experiments were conducted on a small database, future work has to explore whether the results remain stable on a larger set. Additionally, further analysis needs to be performed to answer the question of why the concept-based evaluation measures for ranking systems in a benchmark scenario do not reflect the differences in annotation quality to a large extent, as the kappa statistics or the OS do.

Part III.

Novel evaluation methodologies for multi-label annotation evaluation

Outline

The third part of the thesis presents my contributions on performance measures for multi-label image annotation evaluation. It includes the definition of a fine-grained evaluation measure that cares about semantic relatedness of indexed terms by utilising ontologies and semantic relatedness measures. The measure is developed along the requirements of a new user model on image retrieval and annotation. I perform a study that relates the newly proposed measure to traditional evaluation measures. The strengths and weaknesses of traditional evaluation measures for the evaluation of multi-label annotation systems are determined and outlined. Further, the measure is extensively compared to different ways of assessing the semantic relatedness between terms, including knowledge sources such as WordNet, Wikipedia, Web-search engines, and Flickr, and the qualities of the measure are evaluated regarding stability and ranking performance. This analysis is accompanied by a user study and the proposal of a new dataset on semantic relatedness among visual concepts. Finally, a second evaluation measure is proposed. In contrast, it relies on ranked predictions of the annotation systems, but follows all other user requirements. The measure is compared to other measures for ranked predictions regarding system ranking performance and stability. As baseline for the experimental work in this part, the ImageCLEF 2009 and 2010 submissions to the Photo Annotation task are utilised. Finally, the contributions on test collections and performance measures are united in the chapter about the Photo Annotation task in ImageCLEF. This work analyses and discusses the 180 submissions from three ImageCLEF cycles which constitute a representative selection of the ability of state-of-the-art image indexing systems.

7. A fine-grained evaluation measure for multi-label annotation evaluation



In this chapter, a novel, fine-grained evaluation measure for the evaluation of multi-label annotation approaches is introduced. It directly tackles the score prediction dimension and determines concept-dependent costs based on information that is encoded in an ontology. The measure uses relationships such as disjoints or hierarchical constraints to decide how meaningful and realistic a set of annotations is. Depending on the closeness between the predicted and the ground truth label set, the degree of misclassification is assessed and represented in the overall score. Further, human annotator agreement on concepts is incorporated as weighting factor. The chapter is structured as follows. After the motivation, issues in formulating a new performance measure for annotation evaluation are analysed and discussed in Section 7.2. This includes a user model for media annotation evaluation and requirements on hierarchical multi-label performance measures. In Section 7.3, the performance measure for fine-grained multi-label annotation evaluation, the OS, is defined and exemplarily compared to hierarchical performance measures in Section 7.4. Section 7.5 presents two case studies that relate the OS to 12 other established evaluation measures for annotation evaluation. Their behaviour is compared on the ImageCLEF 2009 submissions and analysed with respect to both random annotations and over-annotation. Finally, the chapter concludes with a summary in Section 7.6. Part of the work presented in this chapter has been published in Nowak and Lukashevich (2009) and Nowak et al. (2010).

7.1. Motivation

Traditionally, research in machine learning has been focused on defining single-label classifiers where labels were mutually exclusive by definition. Applied to image classification, this means that a photo depicts, for example, either a landscape, a city, or persons. However, many real-world situations require classes that are not mutually exclusive. Mostly, the documents contain

aspects of different categories and should be labelled with all relevant items. In the same way as standard single-label classifier approaches cannot be applied straightaway to multi-label classification, multi-label classification requires the definition of its own evaluation measures. In the past, researchers have proposed several measures for *concept-based* and *example-based* evaluation as introduced in Chapter 3.3. These measures can be classified into the three dimensions *measurement direction*, *prediction format*, and *relevance format*, as explained in Chapter 4.3. However, most measures are directly adopted from uni-label classification evaluation and consider each concept in isolation. They therefore ignore the *score prediction* dimension and simply assign binary scores in effectiveness measurement. Further, the measures employed in uni-label classification follow a different user model and lack a connection to a user model for annotation evaluation.

7.2. User requirements for a new multi-label evaluation measure

In the following, a user model for concept-based search and annotation evaluation is proposed and requirements on performance measures are summarised.

7.2.1. User model for annotation evaluation

A main problem of experimental evaluation in IR is caused by the fact that the user is often left out of the evaluation loop. Despite the relevance assessment process, the evaluation of systems is mostly performed independently of any user interaction. With the definition of user models, it is attempted to put the user intentions back into the evaluation setup and to find metrics that reflect these user intentions. While several user models have been identified for IR and browsing (Dupret and Piwowarski (2010) and de Vries et al. (2004)), no user model for indexing based retrieval is in use. Instead, indexing approaches are usually evaluated with concept-based evaluation measures in a single-concept retrieval scenario (see Table 4.2).

Let us consider a search scenario for images. The user requests outdoor pictures that depict a tree in front of a sunset. Consequently, the search engine searches for images labelled with `outdoor`, `tree`, and `sunset`. The requester expects images that show all three visual concepts to be retrieved while being less interested in all outdoor images depicting other landscape scenery. This is even aggravated if the search complies with a professional need to find fitting visual documents, as it may occur in advertising or journalism. The performance evaluation should follow the expectations and intentions of the user and judge the system according to its ability to label images with a set of concepts, instead of evaluating each concept in isolation. This observation leads to the first requirement for annotation evaluation:

- *The performance assessment should consider the complete set of ground truth labels.*

Second, a predicted label set may not contain exactly the concepts requested because of misclassifications or a small database. In this case, the user may prefer near matches instead of unrelated matches or no matches at all. To give an example, a request for images depicting the concepts `{sea, sky, sunset}` is better answered with an image depicting the concepts `{river, sky, sunset}` than with one illustrating `{sea, sky, people}` as can be seen in Figure 7.1. This is due to the semantic relatedness between `river` and `sea` in contrast to `people` and `sunset`. Sometimes, a river and the sea even cannot be visually distinguished in images. To address this issue, the performance evaluation should not only judge misclassifications; rather it has to find a way of distinguishing between sets with an equal number of misclassifications. This lets us derive the second requirement:

- *The measure should judge misclassifications in a fine-grained fashion.*



Figure 7.1.: Images illustrating the example for retrieving *river, sky, sunset* and *sea, sky, people* instead of *sea, sky, sunset*. Photos from Flickr taken by Radio.Guy and tillwe.

A performance measure for multi-label annotation evaluation should consider the user model and inherit both requirements. In the terminology of dimensions of evaluation measures, this means that the measure should follow the horizontal measurement dimension and adopt an example-based measure, as well as include an adaptive or even semantic-based approach to determine the relatedness of misclassifications in the score prediction dimension.

7.2.2. Requirements on multi-label evaluation measures

In addition to a user model, the definition of performance measures should adhere to several requirements and desired properties as illustrated in Chapter 4.6. These rather general requirements can be refined if the evaluation measure is granted access to relationships between concepts in the form of a hierarchy. Kiritchenko (2005) lists requirements for hierarchical evaluation measures. The measures should consider partially correct classification. This means that if a misclassified label is in the same subtree as the true label, the misclassification should be penalised less than if the true label is in another subtree. Second, the misclassification error should increase with growing distance in the hierarchy, and third, the error should be higher in case of misclassifications at the top of the tree than at the bottom. These requirements correlate with the need for judging misclassifications in a fine-grained fashion, as was detailed as second requirement in the user model. However, Kiritchenko explicitly restricts this issue to hierarchies that are used to determine the semantics. The structure of a hierarchy, however, is only one means that can be used to determine the relatedness of concepts. Additionally, Kiritchenko states the need for simplicity and generability of measures.

The definition of a multi-label classification evaluation measure has to solve two further issues in addition to the generally required properties:

1. How to determine the contributions for misclassifications to the overall score?
2. How to map the predicted label set to the ground truth label set?

7.3. **Ontology Score: A fine-grained multi-label classification performance measure**

Ontologies, hierarchies, or taxonomies can provide knowledge that is useful in an annotation evaluation scenario. They allow to determine the *degree of misclassification*. An image which is classified to contain *water* instead of *sea* was not incorrectly classified, but the classifier assigned a more general concept than the expected detailed one (compare Figure 7.2). If the classifier

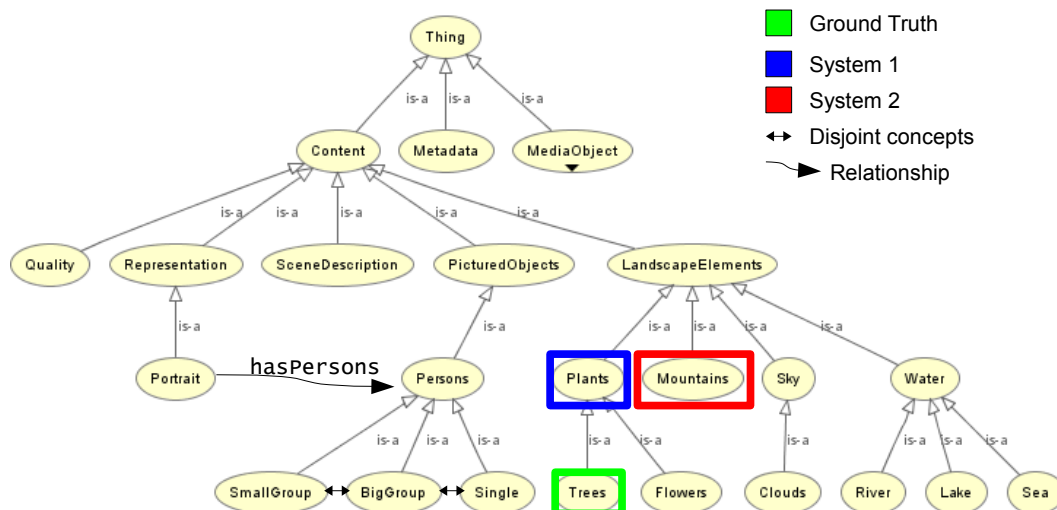


Figure 7.2.: Visualisation of a fragment of the ontology for image annotation. The concepts are hierarchical structured and different types of relationships are exemplarily highlighted.

assigned `sea` instead of `river`, it mistakenly chose a sibling concept, but the classification is not as wrong as it would be if it had assigned the concept `trees`. Further semantic knowledge is provided if the concepts are organised in an ontology. Then, next to the *is-a* relationship of the hierarchical organisation of concepts, other additional relationships between concepts determine possible label assignments. The ontology restricts, for instance, that for a certain sub-node only one concept can be assigned at a time (disjoint items), or that a special concept (like `portrait`) postulates other concepts such as `persons`.

As a result of the preceding analysis, the need for example-based multi-label annotation evaluation measures that adhere to a user model and include semantics in judging relevance becomes apparent. In the following, I propose a novel measure for multi-label classification evaluation called OS. The OS is formulated as an example-based evaluation measure and is bound to an ontology. It overcomes the limitations of traditional evaluation measures that cannot differentiate between semantically false annotated labels and labels with a meaning similar to the correct label. The OS considers three characteristics to judge misclassifications in a fine-grained fashion:

1. Depth-dependent distance-based misclassification costs between concepts

It incorporates the DDMC explained in Chapter 4.4.1 for the uni-label classification case. In this work, it is extended to the multi-label classification case. The costs between each pair of concepts is determined dependent on their distance and depth in the hierarchy and saved in a costmap. The vocabulary of concepts is fixed from the beginning in an annotation task which allows pre-computing the costmap. The calculation of misclassification costs favours systems that annotate a photo with concepts close to the correct ones, in contrast to systems that annotate concepts that are more distant in the hierarchy.

2. Ontology-based penalty

Next to the *is-a* relationship in the ontology, different relationships between concepts are defined. The ontology restricts, for instance, that for a certain sub-node only one concept can be assigned at a time (*disjointness*), or that concepts postulate other concepts. The relations in the ontology are used to verify the co-occurrence of labels for one image. If the system prediction violates relationships, a penalty is assigned instead of applying the depth-dependent costs.

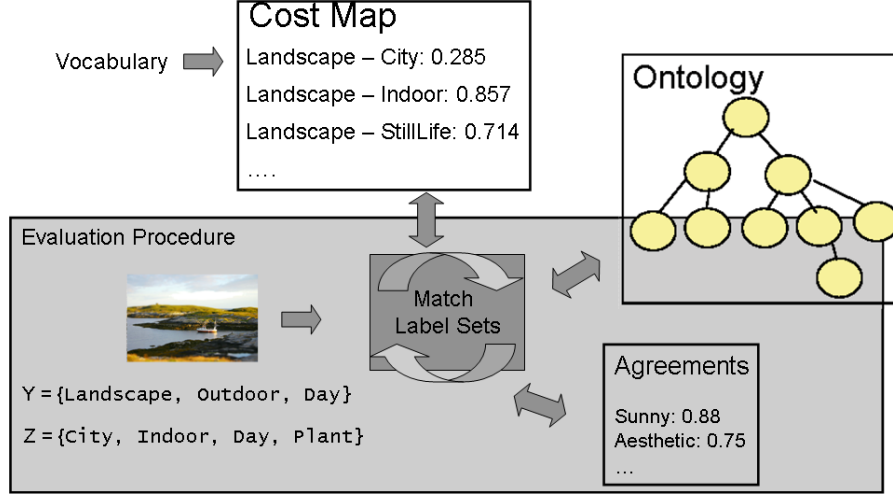


Figure 7.3.: Schematic representation of the components of the OS.

3. Annotator agreements

The OS takes the subjectivity inherent in determining a ground truth into account. It considers empirically determined inter-annotator agreement to rescale the costs. In a user study, 11 annotators annotated all concepts of the ontology in a validation set. The annotation of the majority of annotators was regarded as correct and the percentage of annotators that annotated correctly equals the agreement factor (see Nowak and Dunker (2009a)). The outcome of the user study is an agreement-map. The annotator agreements are a measure for the objectivity of judgements. The greater the disagreement on a concept computed over the validation set, the lower the factor. A perfect agreement results in a factor of one.

The second issue deals with how to map the ground truth set to the predicted label set. In example-based evaluation, the overlap of the system prediction and the ground truth are, in most cases, partly correct. In the OS, a matching procedure maps labels from the system annotations to the ground truth and vice versa in a cost minimisation procedure.

First, the false positive labels $\mathcal{Z}'_n = \mathcal{Z}_n \setminus (\mathcal{Z}_n \cap \mathcal{Y}_n)$ and the missed labels $\mathcal{Y}'_n = \mathcal{Y}_n \setminus (\mathcal{Z}_n \cap \mathcal{Y}_n)$ are computed, as only a match between these labels is necessary. Please note that $|\mathcal{Z}'_n| + |\mathcal{Y}'_n| \leq |\mathcal{Z}_n \cup \mathcal{Y}_n|$ is always valid, because the number of false positive and missed labels can never be greater than the number of the union of labels in both sets. If $\mathcal{Z}_n = \emptyset$, the matching costs for all labels of $\mathcal{Y}'_n = \mathcal{Y}_n$ are set to the maximum. A crosscheck on the predicted label set \mathcal{Z}_n is performed. If labels in \mathcal{Z}_n violate relationships from the ontology (e.g., by assigning labels such as `indoor` and `outdoor` to one photo at the same time), these labels are assigned the maximum costs of one as penalty and are removed from \mathcal{Z}'_n , \mathcal{Y}_n , and \mathcal{Y}'_n if contained. This ensures that the measure does not assign costs twice.

Next, for each concept c_i from \mathcal{Z}'_n a match to a concept c_j from \mathcal{Y}_n is calculated, and for each concept c_h from \mathcal{Y}'_n a mapping to a concept c_g from \mathcal{Z}_n is performed in an optimisation procedure that determines the lowest costs between two labels:

$$\text{match}(\mathcal{Z}_n, \mathcal{Y}_n) = \sum_{c_i \in \mathcal{Z}'_n} \left(\left(\min_{c_j \in \mathcal{Y}_n} \text{cost}(c_i, c_j) \right) \cdot a(c_j^*) \right) + \sum_{c_h \in \mathcal{Y}'_n} \left(\left(\min_{c_g \in \mathcal{Z}_n} \text{cost}(c_g, c_h) \right) \cdot a(c_h) \right), \quad (7.1)$$

with $c_j^* = \text{argmin}_{c_j \in \mathcal{Y}_n} (\text{cost}(c_i, c_j))$. $a(c)$ is the annotator agreement factor for a concept c and ranges between $[0, 1]$. The cost function $\text{cost}(c_i, c_j)$ depends on the shortest path in the hierarchy

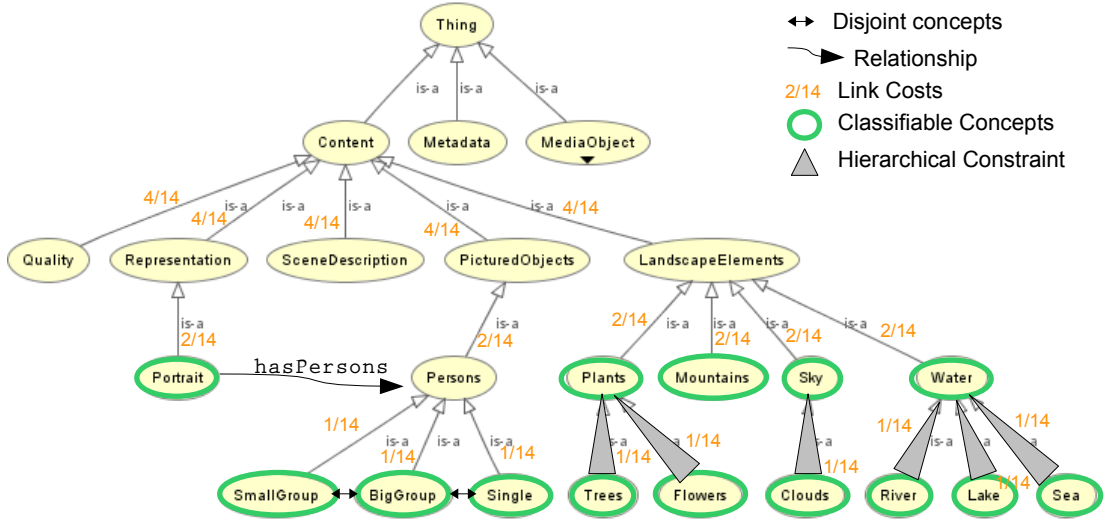


Figure 7.4.: Fragment of the PTO that is used in the example. All concepts that can potentially be assigned to an image are marked with green borders. Further, all ontology relations and the costs for each link are depicted.

between two concepts c_i and c_j . Each link in the hierarchy is associated with a cost that is cut in half for each deeper level of the tree and that is maximally equal to one for a path between two leaf nodes of the deepest level. The costs for a link at level l of the hierarchy are calculated as follows:

$$\text{cost link}_l = \frac{2^{(l-1)}}{2^{(L+1)} - 2}, \quad (7.2)$$

L being the number of links from the leaf node to the root. Finally, the costs $\text{cost}(c_i, c_j)$ are calculated by summing up all link costs at the shortest path between these concepts.

The overall score for the OS is then determined as follows:

$$OS(Z, Y) = \frac{1}{N} \sum_{n=1}^N \left(1 - \frac{\text{match}(Z_n, Y_n)}{|Z_n \cup Y_n|} \right)^\alpha. \quad (7.3)$$

The score is equal to one if all concepts are correctly annotated and goes to zero if no concept was found. Additionally, Shens α -factor, ($\alpha \geq 0$), introduced in Chapter 3.3.4, weighs the strictness of the score regarding fully and partly correct annotations, depending on the application demands.

To summarise, the OS is constructed of the three components cost calculation, ontology relationships, and annotator agreements, which are connected in the matching procedure as schematically presented in Figure 7.3. The measure is still applicable if one or more of these components are missing. The missing annotator agreements can be replaced with an equal factor of 1 for each concept. In case no ontology is present, the component penalising the score computation for knowledge violations can be ignored. The cost computation is linked to the hierarchy of the ontology in this proposed version. However, other means of obtaining the relatedness between concepts can replace this cost computation, as will be presented in Chapter 8.

Table 7.1.: This table shows the score achieved with different hierarchical performance measures for the exemplary comparison illustrated in Section 7.4. The third column denotes the best score that can be achieved with the particular measure. The abbreviations hP, hR and hF stand for hierarchical precision, hierarchical recall and hierarchical f-measure.

Approach	Score	Best Score	Remarks
h-loss (Cesa-Bianchi et al. (2006))	0.29	0	added normalisation
Maximal Loss (Cai and Hofmann (2007))	0.29	0	
weighted Euclidean distance (Struyf et al. (2005))	1.72	0	
hierarchical F-measure (Kiritchenko et al. (2005))	0.66	1	hP = hR = hF = 0.66
hierarchical F-measure (Verspoor et al. (2006))	0.66	1	hP = hR = hF = 0.66
hierarchical F-measure (Ipeirotis et al. (2001))	0.66	1	hP = 1, hR = 0.33
OS	0.45	1	
HS	0.84	1	
Precision _{eb}	0.50	1	
Recall _{eb}	0.33	1	
F-measure _{eb}	0.42	1	

7.4. Exemplary comparison of OS to hierarchical measures

This section compares the OS to the introduced hierarchical performance measures. The behaviour of the measures is demonstrated on the example of one image. None of the hierarchical measures is established in image annotation research, as shown in Chapter 4.4. Therefore, only a rudimentary comparison is performed instead of a deep case study. Afterwards, the hierarchical performance measures are analysed according to their applicability to the user model.

The example refers to the annotation of an image with a maximum of 14 concepts using the hierarchy depicted in Figure 7.4, which is part of the PTO. The concepts with the green border mark concepts that can be potentially assigned to the image. All other concepts are used to structure the concept hierarchy and cannot be directly assigned. The figure shows the relationships that are used in the OS, such as the hierarchical constraints, disjoints, and free relations. Each link is associated with a cost that was computed based on the formula in Equation 7.2.

The example considers the labels {portrait, single_person, plants, water, mountains, sky} as ground truth set. In contrast, the system predicted the labels {portrait, trees, mountains, clouds}. These annotations are graphically presented in Figure 7.5 a) as black nodes. Table 7.1 lists the scores that are assigned in evaluation by different hierarchical performance measures. Only those measures that can be directly computed with the hierarchy information are depicted. Measures that need ranking information, statistics from a training set, or other additional information are not considered. In particular, the measures h-loss, Maximal Loss, weighted Euclidean distance, three variants of the hierarchical F-measure, OS, Hierarchical Score (HS), and the common example-based Precision, Recall, and F-measure are analysed.

The two loss functions h-loss and Maximal Loss assign the same score. The h-loss does not consider errors for descendants in the hierarchy in cases where an error was already counted in an ancestor. It therefore counts the false negative label plants, but ignores the false positive label trees. Similarly, it counts the error for the false negative sky and ignores the false positive clouds. A normalisation to the proposed measure is added by dividing by the number of concepts

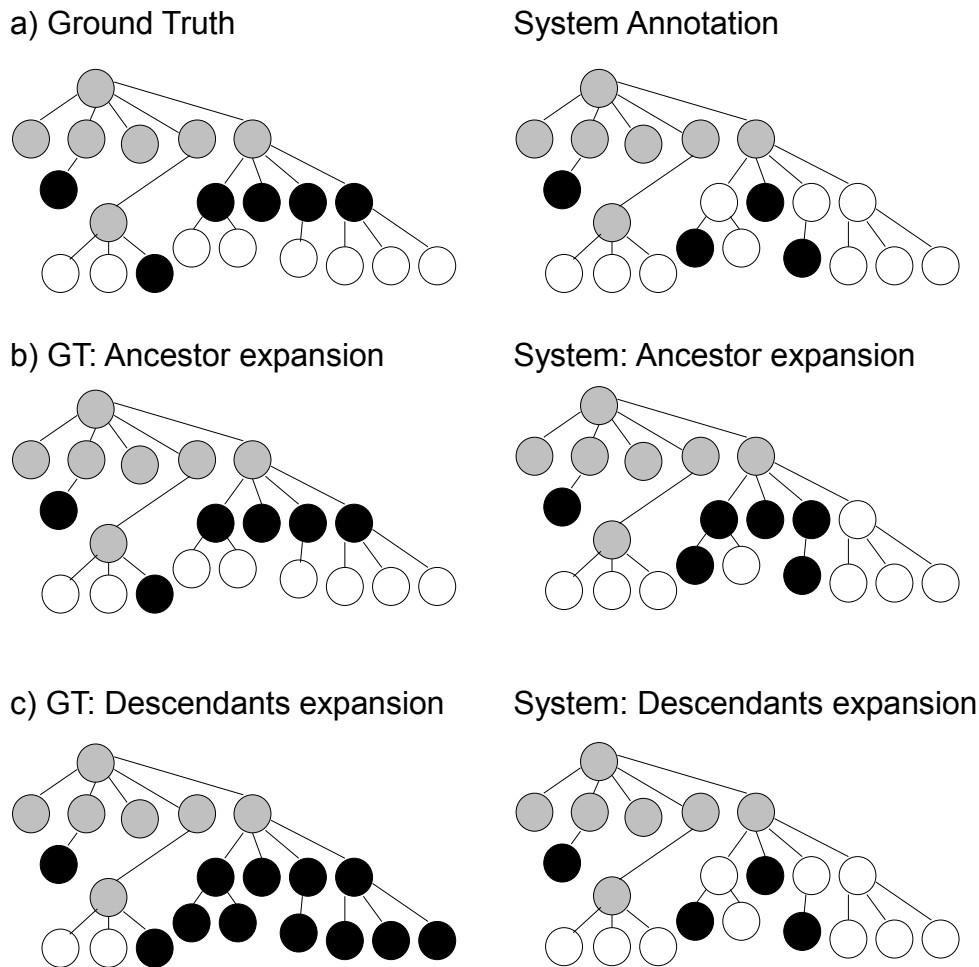


Figure 7.5.: Schematic representation of the PTO fragment of Figure 7.4 that is considered in the example. The grey nodes denote concepts that cannot be used in the annotation, black nodes denote actually annotated concepts, and white nodes represent concepts that are not annotated. In Subfigure a), the annotations for the ground truth and the system are depicted. Subfigure b) shows the annotations augmented with the ancestors and Subfigure c) illustrates the annotations extended with the descendants.

(in this case 14), as it is done in the Hamming Loss and Maximal Loss measures. The Maximal Loss considers the annotation sets with expanded ancestors. These annotations are shown in Figure 7.5 b). While no ancestors can be expanded in the ground truth, the system predictions are augmented with the labels `plants` and `sky`.

The weighted Euclidean distance measure calculates the Euclidean distance with weights for each link that are based on the depth in the hierarchy. Its score is not bound to a value between 0 and 1, which makes it very counterintuitive and difficult to interpret.

The hierarchical F-measures result in the same score for all three variants. The different way of averaging between the approach of Kiritchenko and Verspoor is only apparent when more than one image is used in the score computation. The approach by Ipeirotis is based on the augmented sets including the descendants of the annotated labels (see Figure 7.5 c)). This approach results in a hierarchical Precision (hP) of 1 and a hierarchical Recall (hR) of 0.33, but after averaging, it delivers the same result as the other two approaches.

The OS is computed as explained in Section 7.3 with equal annotator agreements of one for

all concepts. The ontology restrictions result in a penalty for ignoring the hierarchical relations in the case of `trees` and `clouds`, and for ignoring the relationship by predicting `portrait` without a person concept. The other costs are obtained by the optimised hierarchical distance between concepts. The HS is also included, which is computed in the same manner as the OS but does not consider the ontology penalisation. The HS measure assigns a performance of 84%, as the distances in the hierarchy for misclassified concepts are short. However, these misclassifications violate real-world knowledge. This cannot be captured by the measure if it is reduced to hierarchical costs. Finally, the table shows the conventional example-based Precision, Recall, and F-measure that are computed solely on the label sets and not linked to the hierarchy.

All presented hierarchical measures relate to the two requirements of the user model. They evaluate based on the set of predicted and ground truth labels, and they assess misclassifications in a fine-grained fashion. Most make use of an ancestor or descendant expansion of the label sets on which the cost calculation is based. However, the measures assume that links in a hierarchy represent uniform distances and that concepts in one subtree of a hierarchy have similar meanings, although they might violate real-world knowledge. All in all, the scores for the different performance measures are difficult to compare. A detailed study of the hierarchical measures would be necessary to assess their effectiveness for image annotation evaluation. However, none of the reported measures have been applied in image annotation evaluation. Therefore, I want to focus on the characteristics of the complete OS measure concerning different properties and compare it to established performance measures for multi-label annotation evaluation in the following.

7.5. The effectiveness of performance measures in image annotation

Two case studies are performed with the goal of analysing and comparing commonly applied performance measures for the evaluation of multi-label annotation approaches. The behaviour of different evaluation measures on the submissions to the ImageCLEF VCDT in 2009 is investigated. The task poses the challenge of the automated annotation of 13,000 Flickr photos with 53 visual concepts. In total, 73 runs were submitted by 19 research groups. A run denotes one configuration of a system in which predictions are assigned to the complete test collection. The participants were asked to submit confidence values in the range of [0:1] for each concept in each photo. While most groups followed this request, some submissions only contained binary predictions. Disadvantages of these binary submissions when evaluated with rank-based performance measures will be highlighted. A threshold of 0.5 was agreed upon for the performance measures that need a binary decision in order to judge presence or absence of a concept. Each group was allowed to submit up to five configurations. In the official benchmark, the submissions were evaluated with the performance measures EER, AUC, and OS. Details on the task can be found in Chapter 10, as well as in Nowak and Dunker (2010).

7.5.1. Choice of configurations for the case study

In the case studies on the characteristics of evaluation measures in image annotation, not all submitted configurations are always considered, due to a more intuitive visualisation of the results. In some plots, only one configuration per group is utilised. For this, not the best submitted configuration of every group, but the configuration with the largest variance between the result ranks for EER and OS, is chosen. In Table 7.2, part of the official results for the EER and the OS measure are displayed. The systems of the participants are numbered from 1 to 19 in alphabetical order. As the goal is to analyse the differences and weaknesses of the evaluation measures, the author believes that these 19 configurations are the most interesting ones for the case study. Further, it is exemplarily confirmed that the characteristics of evaluation measures

Table 7.2.: The table shows the configurations with the largest variances in the ranking of the official VCDT results for the measures EER and OS. One configuration of each participating group was selected.

System	Rank	EER	Rank	OS	Rank Diff
System 9	1	0.234	21	0.740	-20
System 11	5	0.250	23	0.731	-18
System 5	11	0.256	2	0.810	+9
System 19	14	0.267	1	0.811	+13
System 6	15	0.272	9	0.793	+6
System 3	17	0.292	40	0.613	-23
System 13	24	0.331	53	0.482	-29
System 7	26	0.342	67	0.368	-41
System 14	31	0.357	66	0.376	-35
System 12	35	0.384	72	0.261	-37
System 15	40	0.440	28	0.716	+12
System 18	46	0.452	11	0.779	+35
System 2	48	0.454	62	0.396	-14
System 10	53	0.467	10	0.786	+43
System 17	54	0.479	32	0.690	+22
System 1	56	0.483	17	0.756	+39
System 8	59	0.486	20	0.744	+39
Random 0	-	0.500	-	0.384	-
System 4	70	0.502	29	0.709	+41
System 16	73	0.527	65	0.385	+8

are valid for all submitted runs. Additionally, several random configurations are investigated. The configuration Random0 stands for the results of uniformly distributed pseudo random numbers that varied between [0:1]. All other random runs are denoted by RandomXX, where XX stands for the percentage of 1 values in the annotations. Table 7.2 clearly shows significant differences in the rankings for the evaluation with the concept-based EER that considers the ranked predictions and the example-based OS that relies on the predicted binary set. Part of these differences may result from the post processing of the systems and the thresholding of confidence values.

7.5.2. Evaluation measures

The case studies compare 13 commonly applied concept-based and example-based evaluation measures in image annotation evaluation. In particular, the concept-based binary measures Precision, Recall, and F-Measure and the concept-based measures for ranked annotations AUC, EER, as well as iAP are analysed. They are compared to the example-based measures for binary predictions Precision, Recall, F-Measure, Accuracy, Alpha Evaluation, HS, and OS, as listed in Table 7.3.

7.5.3. Study 1: Influence of overannotation on performance score

The first case study investigates the influence of the LD on the result score. It defines how many concepts are assigned to each photo on average, divided by the total number of available

Table 7.3.: Performance measures that were utilised in the case studies categorised by their type.

	concept-based measures	example-based measures
binary predictions	Precision Recall F-Measure	Precision Recall F-Measure Accuracy Alpha Evaluation HS OS
ranked predictions	AUC EER iAP	

concepts. As the LD is calculated based on binary predictions, only the 10 performance measures that consider binary predictions are analysed. The alpha evaluation measure is calculated for the values $\alpha = 0.5$ and $\alpha = 0.2$. In the case of $\alpha = 1$, it equals the example-based Accuracy.

The ground truth annotations of the test set show a LC of 9.06 and a LD of 0.17. This means that, on average, nine concepts per image were regarded as relevant by the human judges. Figure 7.6 depicts the LD plotted against the results of the measures for all 73 runs as dots. The crosses represent the random configurations and the star depicts the score of the ground truth. The indices attached to each symbol denote the name of the run.

Subfigure a), b), and c), show the concept-based Precision, Recall, and F-Measure, while Subfigure d), e), and f) present the results of their example-based variants. Both recall measures result in increased scores for an overannotation of images. This behaviour is inherently captured in the measure which evaluates how many of the actual annotations have been found. Consequently, a prediction containing all concepts results in perfect Recall. Four submissions followed this example. In contrast, the Precision is not negatively influenced by the LD. Results are likely to score better if the LD is close to the LD of the ground truth. However, it is not sufficient to only predict the number of concepts that are close to LD; rather the concepts have to be correct, as can be seen from the random configurations. While, in general, the example-based variants assign higher scores, these are better able to discriminate between random configurations and system predictions. The variants of the F-Measure are determined by averaging Precision and Recall. Therefore, they inherit the tendency to assign better scores for a higher LD. However, as can be seen in the plots, this influence is not fundamental. The random runs with a high percentage of annotations obtain only slightly better scores than the ones with a lower percentage. Overall, the concept-based variant shows a low discriminancy between system and random predictions.

The Subfigures g), h), and i) illustrate the results for the alpha evaluation for $\alpha = 1$, $\alpha = 0.5$, and $\alpha = 0.2$. With decreasing alpha, the measure is more forgiving and assigns higher scores. Also, the random configurations score higher, especially for configurations with a percentage greater 20% annotated concepts. While the measure can differentiate well between random configurations and submissions for $\alpha = 1$, this discriminancy degrades for lower alpha values. Nonetheless, an overannotation does not lead to high scores, but submissions with a LD around that of the ground truth score highest. However, this discriminancy also degrades with decreasing alpha.

In the Subfigures j), k), and l), the scores for the iAP, HS, and OS are presented. The scores for the iAP are independent of the LD, which is shown by the random configurations. As for the

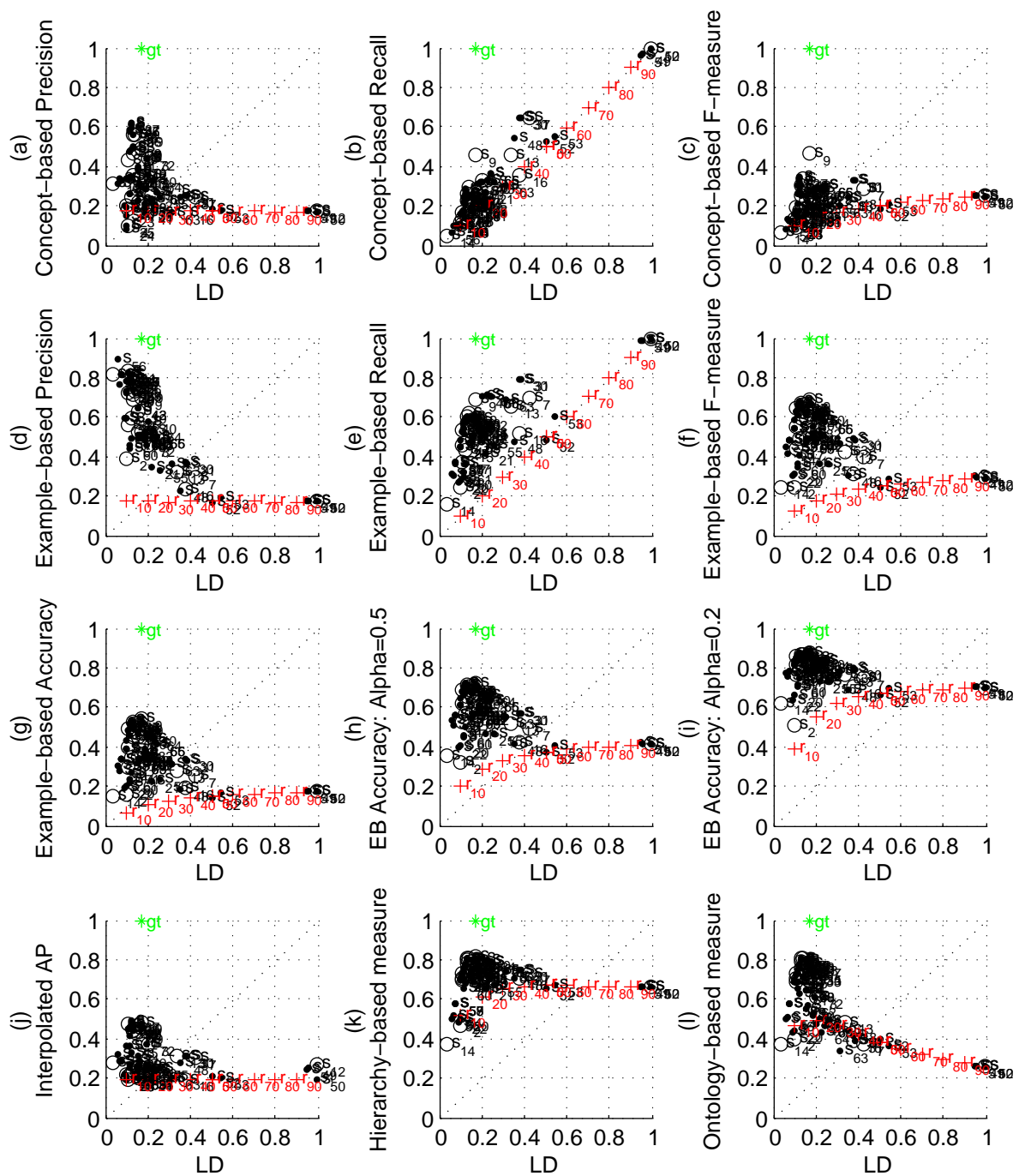


Figure 7.6.: Scatter plots on the influence of the label density to the score of different performance measures. The plots depict the 73 runs of the participants in black and 10 random configurations in red. The score of the ground truth is highlighted in green.

variants of precision, the score is not changed for changing label densities. Similarly to the alpha evaluation with $\alpha = 0.2$, the HS assigns higher scores for random configurations with more than 20% annotated concepts. In both measures, a complete annotation of all concepts generates scores in the median of all system scores. The discriminancy between random configurations and system annotations is therefore not given. The OS overcomes the limitations of the HS. Although it tends to assign good results to configurations with similar LD as the ground truth, the discriminancy between random configurations and good system scores is much higher. An overannotation of the photos does not lead to good results in the OS. The best results are assigned to runs that annotated close to the LD of the ground truth. However, the random runs with a density similar to the density of the ground truth are assigned higher scores.

7.5.4. Study 2: General characteristics of example- and concept-based measures

The second case study concentrates on general characteristics of the performance measures. First, the selected run configurations that were chosen based on the variance in ranking between the concept-based EER and the example-based OS are analysed in detail (see Table 7.2). The focus is placed on variances in scores of individual runs for different measures. Second, the scores for all submissions are visualised for pairs of measures. With this second analysis, it is proven that the characteristics of the evaluation measures are valid for all runs. The study considers all performance measures that were introduced in Table 7.3.

7.5.4.1. Results for the selected run configurations

Figure 7.7 illustrates the results of the evaluation of the chosen run configurations. The first row depicts the results for concept-based Precision, Recall, and F-measure. Contrasting row (b) shows the results for the example-based variants of Precision, Recall, and F-measure. Row (c) depicts the scores of AUC, EER, and iAP. For an easier comparison, the results of 1-EER are visualised. Row (d) shows the scores for the α -evaluation measure with different values for α , and finally, row (e) presents the example-based Accuracy, HS, and OS scores. In each bar diagram, the same order of runs is utilised, beginning with systems 1-19, followed by the random runs and the ground truth.

Concept-based Precision, Recall, and F-Measure: Figure 7.7 (a) depicts the results for the concept-based Precision, Recall, and F-measure. The Precision for the submissions varies between 0.1 and 0.6, with an average of 0.3. The random runs achieve a Precision of 0.17. In terms of Recall, the submissions score at minimum 0.05 and at maximum 0.99, with a mean of 0.3. System 12 achieves a Recall of almost 1. This means that nearly all concepts were annotated as present for all images. This fact is also illustrated by the LD, which is 0.99 for System 12. The reason for this behaviour can be two-fold. Either, the system really annotates all concepts as present, or the threshold of 0.5 which is used to map the confidence values to a binary decision is not well selected. Depending on the number of annotated labels, the random runs are assigned a maximum score of 0.9 Recall if 90% of the annotations are set to 1. The F-measure varies between 0.07 (System 14) and 0.47 (System 9) with a mean of 0.22, while the random runs score at maximum 0.24. These values indicate that, with random runs containing a high percentage of annotated concepts, a score higher than the average of the submissions can be achieved. This also holds when regarding all 73 submissions. The mean of all submissions in terms of F-measure is 0.20.

Example-based Precision, Recall, and F-Measure: In Figure 7.7 (b), the results of the example-based Precision, Recall, and F-measure are depicted. Compared to the concept-based ones, the example-based evaluation measures report higher scores, e.g., for the systems 4, 8, and 14. The

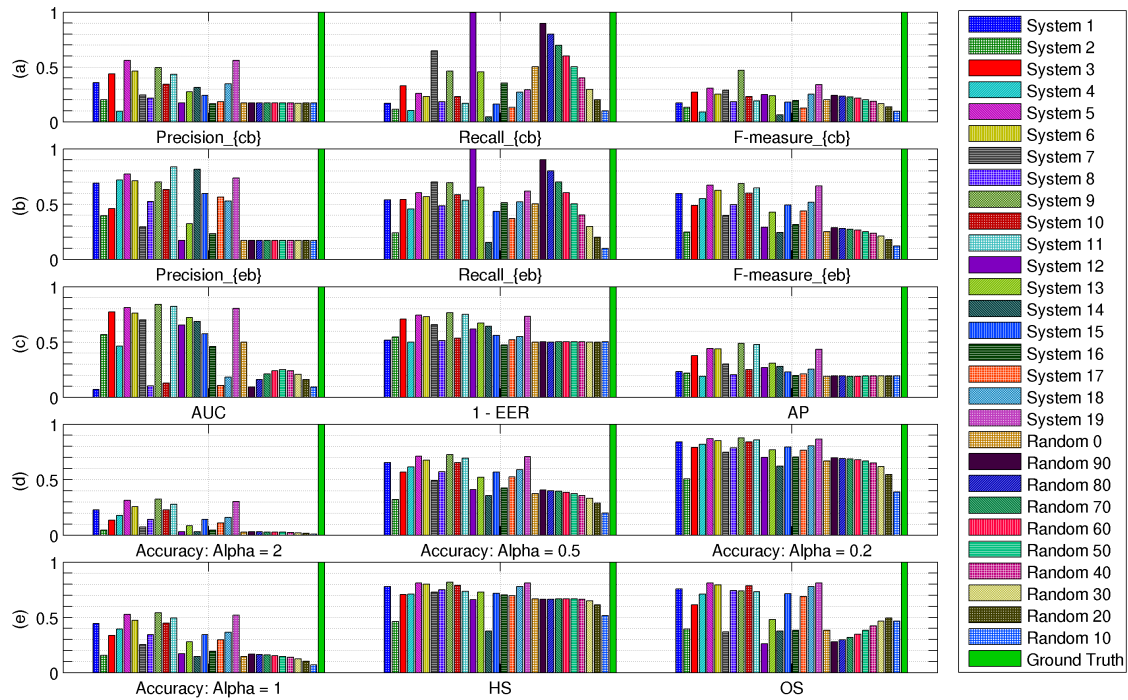


Figure 7.7.: Results of the concept-based and example-based evaluation measures for the chosen run configurations. The runs in the diagrams are ordered ascending to the system numbers, followed by the random runs and the ground truth result for each measure.

average values are 0.56 Precision, 0.54 Recall, and 0.49 F-measure. For the random configurations, one can detect little difference between the example-based and concept-based Precision measure. For both evaluation paradigms, the random configurations do not achieve better results than 0.18 Precision. Taking a look at the plots in row (a) and (b), especially at the example-based and concept-based F-measures, it is interesting that these measures differ not only in scale, but also in the order of the systems. This finding can be explained by different averaging methods. The concept-based F-measure uses averaging over all concepts and all concepts are weighted equally, so that several badly estimated concepts could drastically lower the average score. Choosing the concept-based F-measure, the systems and the random runs are mixed all over the result list. For the example-based F-measure, the random runs are grouped at the end of the result list with the lowest scores, except for System 14. Therefore, it can be derived that using the example-based evaluation, a simple adaption of thresholds of randomly generated scores cannot achieve highly ranked results, which is an indicator for manipulation possibilities in the concept-based F-measure.

Concept-based AUC, EER, and iAP: The results for the concept-based measures AUC, EER, and iAP are presented in row (c). These measures are calculated based on ranked annotation results and use the confidence values provided by the participants. The results show that in terms of AUC, scores between 0.07 and 0.84 were achieved, with a mean of 0.54 for the submissions. The random runs are assigned a maximum score of 0.5 in case of the run with random numbers in the interval $[0:1]$ and worse (< 0.25) in case of binary runs. In terms of 1-EER, the scores range between 0.47 and 0.77, and the random runs obtain a score of 0.5. For iAP, the mean score is 0.31, ranging between 0.19 and 0.49. Randomly, a maximum score of 0.19 can be achieved.

Example-based Alpha Evaluation and Accuracy: Figure 7.7 (d) illustrates the results of the α -evaluation measure with parameter values (2, 0.5, 0.2) for α . The results for $\alpha=1$, which equals

the example-based Accuracy, are depicted in Figure 7.7 (e). It is apparent that with smaller values of α , the results of the systems become better as the measure becomes more forgiving. Despite of System 14, all lower ranks are occupied by the random configurations. For $\alpha = 0.2$, even with random runs, a score of about 0.65 is achieved. It is also obvious that with the α -evaluation, not the whole range of values can be achieved by the systems. E.g., for $\alpha=2$, the best system (apart from the ground truth) achieves a score of 0.32 and the worst one of about 0.01. This is a difference of 0.31. With decreasing α the interval grows over 0.47 ($\alpha = 1$) to 0.53 ($\alpha = 0.5$) and decreases to 0.49 ($\alpha = 0.2$).

Example-based Hierarchical and Ontology Score: The results for HS and OS are illustrated in Figure 7.7 (e). A classification score of about 0.65 can be achieved with the random configurations in terms of HS. These results show that the HS cannot differentiate between good and bad classification systems. The OS measure reports better results. The system scores vary between 0.26 and 0.81, with an average of 0.63 in terms of OS. The random runs achieve a maximum value of 0.49. The OS tends to generate good results if the annotations show a density comparable to the LD of the ground truth. This can be observed e.g., for System 17, which has a low value in terms of concept-based F-measure (0.12), a bad score in terms of AUC (0.11), but an OS of 0.69 and a LD of 0.11. Systems that obtain low scores in the AUC and F-measure can achieve good results in the OS if they adhere to the ontology rules. Good systems in terms of AUC and F-measure remain with good results in the OS.

7.5.4.2. Generalisation of the results to all submissions

The scatter plots in Figure 7.8 visualise the results of all 73 submitted configurations, the 10 random configurations, and the ground truth. Exemplarily, pairs of evaluation measures were chosen and plotted against each other. The runs that were utilised in the bar diagrams of Figure 7.7 are denoted as large circles and the other runs as small circles. Further, the crosses represent the random configurations, while the star depicts the score of the ground truth. The indices attached to each symbol denote the name of the run.

In Figure 7.8 (a), (b), and (c), the example-based Precision, Recall, and F-measure are compared to their concept-based counterparts. The analysis of the results as outlined in the discussion of the bar diagrams can be transferred to all runs. In general, the example-based variants assign higher scores. The concept-based Recall is slightly higher for two systems than the example-based one. For many systems, the assigned labels are uniformly distributed over the examples, while the Recall values for different concepts vary significantly. Due to different averaging methods, this results in a higher example-based Recall, but in a lower concept-based Recall. In the case of the two systems with higher concept-based Recall than example-based, the Recall values are uniformly distributed over the concepts. The random runs achieve the same values for both variants in terms of Precision and Recall. The Precision score is low, but the Recall goes to 1 with rising number of annotations. The random runs are assigned scores that are in the average of all scores in terms of the concept-based F-measure. In contrast, the scores range at the lower bottom of the result lists for the random runs in terms of example-based F-measure.

The characteristics of the iAP measure are compared to the concept-based Precision and 1-EER in the plots (d) and (e), respectively. The scores for the concept-based Precision and iAP are in agreement. The iAP tends to assign stricter scores, which is clearly observable compared to the 1-EER. For both measures, the random runs are assigned lower scores than the systems. Independent of which percentage of annotations are randomly set to 1, the random runs are assigned the same score, which is amongst the lowest score of all submissions. The iAP measure is therefore a stable measure which is robust against manipulations from random runs similar as

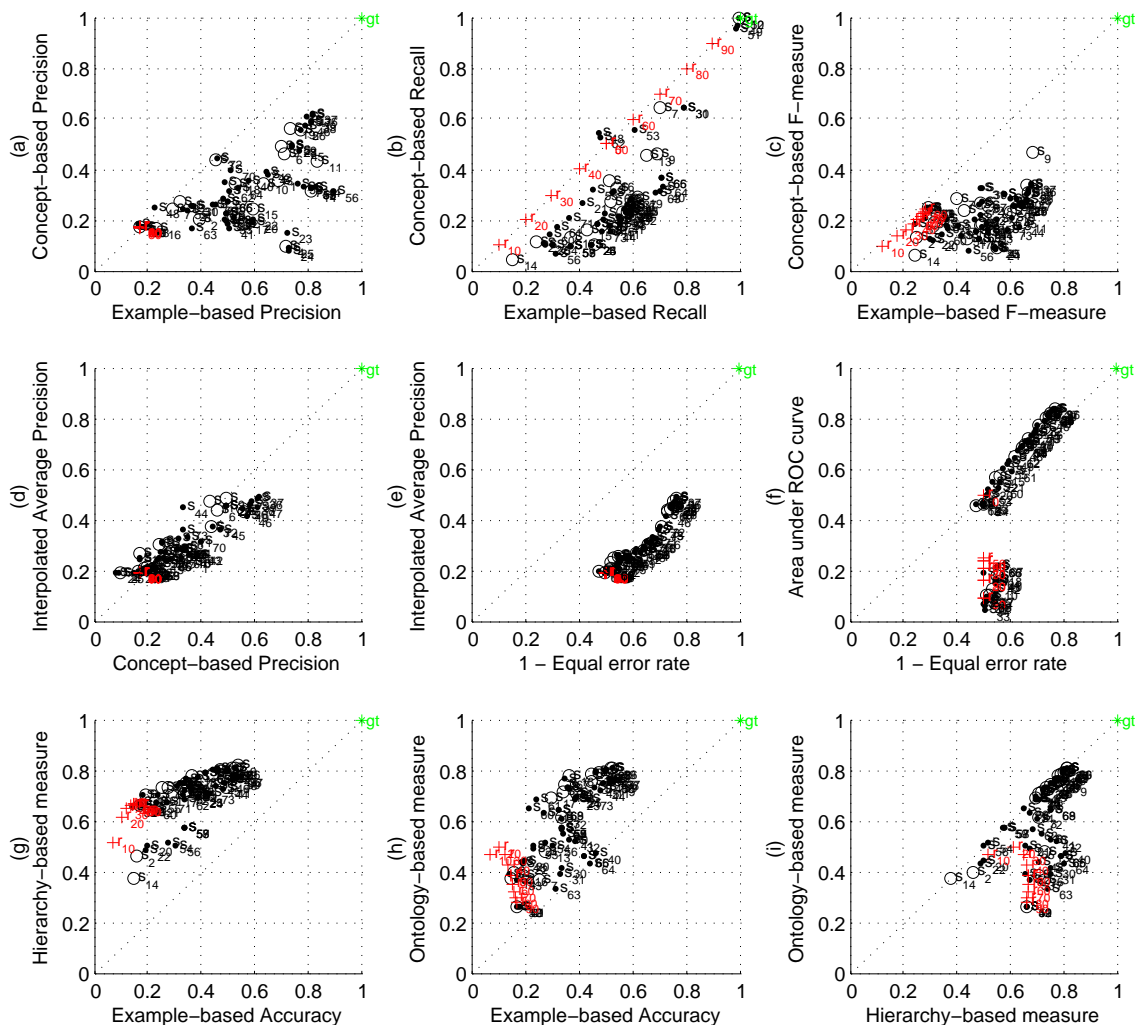


Figure 7.8.: The scatter plots visualise the results for all run configurations for some of the evaluation measures. The submitted run configurations are presented as circles. Big circles denote the runs that were utilised in the bar plots of Figure 7.7 and small circles were utilised for all other submissions. The results for the random runs are presented as crosses and the score of the ground truth as star.

the concept-based and example-based Precision.

Scatter plot 7.8 (f) shows the measure 1-EER compared to AUC. Two clusters are visible. The cluster at top contains the results for systems that submitted confidence values for each annotation. For this, cluster AUC and 1-EER correlate. In some configurations, the annotations were submitted as binary values. The scores for these systems cluster at the bottom of the plot. In terms of EER, the scores between the cluster at the bottom and the worst runs with confidence values are similar. In contrast for the AUC, all runs with binary decisions achieve at most a score of 0.25, and the runs with confidence values obtain at least a score of 0.46. This can, for example, be observed for the run with pseudo random numbers between [0:1], which receives a score of 0.5. Summarising, the AUC measure disadvantages submissions containing binary decisions.

The plots (g), (h), and (i) illustrate the characteristics of the OS measure. As already outlined in the discussion of the bar diagrams, the HS measure cannot differentiate between random runs and submissions of average performance. The OS measure assigns the lowest scores to the random runs. The behaviour of OS and HS is correlated for the top submissions. Keeping in mind that

the difference between both measures are the penalties of the ontology, it is obvious that the top systems are not penalised to a great extent. Therefore, both measures correlate for high performing runs. In contrast to HS and OS, the example-based Accuracy assigns stricter scores.

7.5.5. Discussion

The findings on characteristics of 13 evaluation measures with respect to the LD, to random numbers, and in comparison to each other, are summarised in the following:

- Example-based evaluation: HS and OS

The HS does not satisfy the needs of a good evaluation measure. With random numbers good results can be achieved, and the difference to results of well-working classification systems is not apparent. The OS assigns good scores to systems that got good ranks in other measures, such as the example-based F-measure. However, it tends to give better scores to systems that follow all ontology rules but got only average ranks with traditional measures. Further, the OS assigns better scores if the number of annotated concepts is close to the LD of the ground truth.

- Example-based evaluation: α -evaluation

The results of the α -evaluation are dependent on the threshold chosen for α . If α is equal to 1 or to 0.5, the results for the case study show the best distribution without assigning good scores to random runs.

- Precision, Recall, and F-Measure

The example-based Precision, Recall, and F-measure assign higher scores than their concept-based counterparts. As outlined, this is due to the uniformly distributed annotation quality per item for most systems, in comparison to the varying quality per concept. While the concept-based Precision shows good evaluation characteristics, the concept-based Recall is not adequate as evaluation measure for multi-label evaluation. Further, the study shows that a simple adaption of thresholds of randomly generated scores can achieve high ranked results in an evaluation with the concept-based F-measure. The scores for random runs are not of comparable quality when contrasted to the scores achieved by systems for evaluation with the example-based F-measure. In both cases, there is no major influence on the scores by the number of labels.

- Concept-based evaluation: AUC

AUC clusters the scores for submissions with binary values and submissions with confidence values into two clusters. As a consequence of the definition of the measure, binary submissions can obtain significantly worse results than 0.5, as achieved when random values in the interval [0:1] are used. This leads to the conclusion that the results for binary annotations should not be compared to the ones of confidence predictions. For a benchmark scenario, this includes the strong need for participants to follow the instructions and submit real confidence values when requested. Alternatively, another evaluation measure based on binary predictions should be utilised.

- Concept-based evaluation: Interpolated AP

The scores obtained by iAP show a correlation to the concept-based Precision and the EER, although it assigns stricter scores. Further, iAP is a stable measure, as it is robust against manipulations from random runs and independent of the percentage of annotated labels.

In summary, no preference can be given to either example-based or concept-based measures. Depending on the application, the one or the other may be more appropriate. For example-based evaluation, I suggest using the example-based F-measure, Accuracy, or OS, as, depending on the needs, these measures show good characteristics. The HS should not be utilised in evaluation. Also, the adaptation of the α -evaluation with different parameters cannot convince for α values other than 1. For concept-based evaluation, I recommend using iAP. If it is assured that the evaluation considers solely confidence predictions, the AUC measure shows good characteristics.

7.6. Summary

This chapter introduced a new performance measure for multi-label annotation evaluation that adheres to an ontology. The fine-grained classification costs are derived from the hierarchy and the relations of the ontology, and allow to determine the degree of misclassification for partially correct label sets. A user model for annotation evaluation has been proposed and requirements have been determined from this model. The OS is formulated as example-based evaluation measure to address the requirement on evaluating sets of annotations. Second, it judges misclassifications in a fine-grained fashion by the three components cost calculation, ontology relationships, and annotator agreements. The definition of the OS addresses general requirements on performance measures. In particular, the three requirements of King (2003) concerning the best and worst scores, the monotony of the scores, and the requirement about its automation are addressed by the OS. Further, the requirement of a cheap set up is fulfilled in case an ontology is available. It is not necessary to tune any parameters in order to derive the OS. However, the setup of a new ontology can take some time. The three requirements of Kiritchenko (2005) are all addressed by the depth-dependent misclassification costs. A small example was used to illustrate the behaviour of the OS in comparison to other hierarchical evaluation measures.

In the second part of the chapter, two case studies were performed in order to analyse the characteristics and the behaviour of the OS. The first study analyses the influence of the label density on the final score. In the second case study, the behaviour of the OS is compared to 12 other established performance measures for multi-label annotation evaluation and their strengths and weaknesses are pointed out. Concluding, the example-based F-measure, Accuracy, or OS showed promising results in terms of example-based evaluation. In contrast, the HS cannot cope with traditional evaluation measures. Also, the adaptation of the α -evaluation with different parameters cannot convince for α values other than 1. For concept-based evaluation, iAP is recommended. The AUC measure shows good evaluation characteristics in case all predictions contain confidence values. All presented evaluation measures need a threshold to obtain a binary decision about the presence and absence of concepts for ranked predictions despite AUC, EER, and iAP. These thresholds have a major influence on the results of the evaluation.

In contrast to all standard IR evaluation measures, the OS does not only perform a binary decision when comparing label sets, but calculates scores for each label, also when it is only contained in one of both sets. This evaluation approach seems promising, as concepts annotated semantically close to the correct concept are not regarded as incorrect, but partly correct. The degree of correctness is deducted from the length of the path in the ontology between both concepts. Because of this, it is important to find a commonly agreed upon way of structuring the concepts in the ontology in future work. This could be performed, e.g., with the help of user studies, or by utilising a method based on semantic relatedness to calculate costs. It has to be investigated whether the existing semantic ontology-based measures that were tested on quite different ontologies in size and structure, such as Gene Ontology or WordNet, can be applied to the PTO. Further, the OS would benefit if it is independent from a threshold for confidence-based annotations.

was presented in contrast to traditional evaluation measures in Chapter 7. The OS uses ontology information to detect violations against real-world knowledge in the annotations and calculates costs between misclassified labels. However, there exist cases in which the OS does not work adequately. This occurs because the OS bases its costs computation on measuring the path between concepts in the ontology. Therefore, it assumes that the number of links between two concepts is determined by their mutual similarity. However, links in an ontology do not usually represent uniform distances. These limitations are illustrated on the example of the PTO (see Chapter 5). For instance, the cost obtained between the concepts *Landscape* and *Outdoor* is 0.86, which is the same as between *Landscape* and *Indoor*. However, taking into consideration real-world knowledge, it is more likely that a scene depicting a landscape is an outdoor scene rather than an indoor scene. Another example arises when the cost between *Landscape* and *Trees* yields 0.93. This high value implies that they are quite distant in the ontology; this should mean that the likelihood of them appearing together should be low. However, this assumption contrasts sharply with real-world expectations. Therefore, the computation of this cost is heavily influenced by the structure adopted by the ontology, how dense it is, and by the fact that it is well-balanced or not, rather than by a real semantic distance between terms.

The work on hierarchical evaluation measures as introduced in Chapter 4.4 does not question the appropriateness or ambiguity of the structure of the underlying hierarchy. However, the example on the PTO illustrates how important this issue is in order to fairly judge systems and correctly interpret the evaluation results. In their reviews on hierarchical measures, Costa et al. (2007) and Lord et al. (2003) point to the problem of imbalanced trees and the information gain at leaf nodes for depth-dependent measures. While the information might be of the same quality in two leaf nodes, the question of which level of the tree the leaf node is situated on influences the cost assignment.

To overcome these limitations, the example-based OS measure is extended and the behaviour of different cost functions in the evaluation and ranking of multi-label classification systems is investigated. These cost functions estimate the semantic relatedness between visual concepts considering several knowledge bases, such as Wikipedia, WordNet, Flickr, and the WWW, and allow for a fine-grained evaluation. The goal of this experiment is to analyse the influence of the semantic relatedness measures on the OS and to assess whether their use in computing the cost functions improves performance measure characteristics. In order to achieve this, the OS is applied using different cost functions to a real case scenario, the predicted results (73 runs) provided by 19 research teams during their participation in the Photo Annotation task of the ImageCLEF 2009 benchmarking activity (Nowak and Dunker (2009b)).

8.2. Semantic relatedness measures reconsidered

Related work on the definition of semantic relatedness between two terms was introduced in Chapter 4.5. To summarise, semantic relatedness measures can be categorised into thesaurus-based approaches, Wikipedia-based approaches, and distributional approaches.

This study considers all three approaches to determine the semantic relatedness. It incorporates a total of 9 WordNet measures, including Wu and Palmer (1994) (WUP), Leacock and Chodorow (1998) (LCH), Resnik (1995) (RES), Jiang and Conrath (1997) (JCN), Lin (1998) (LIN), Hirst and St-Onge (1998) (HSO), Lesk (1986) (LESK), Patwardhan (2003) (VEC), and PATH. The measure WIKI (Milne and Witten (2008)) defined on the Wikipedia thesaurus is further employed in the study. Additionally, two distributional measures relying on the WWW are investigated. *Www_G*, uses the Google search functionality and *www_Y* the Yahoo search functionality to determine the semantic relatedness as proposed by Gracia and Mena (2008). In both measures, Equation 4.2 is applied to quantify the semantic relatedness. Finally, the two measures FCS and

FTS that consider the Flickr search results are used (Jiang et al. (2009)). In the study, the FTS and FCS are adapted as relatedness measures for multi-label evaluation. The number of photos on Flickr recently crossed the 4.3 billion threshold, which was used as N in the computation.¹ Measures based on learning of semantic distances, such as the Flickr Distance, are not considered in this work. They rely on visual features of images and a modelling of these features. In a general evaluation scenario, this would include additional parameters and might favour submissions from participants that use the same visual features for the annotation as those used in the evaluation measure.

8.3. Study setup on evaluation using semantic relatedness

Two experiments assess the quality of the semantic relatedness multi-label evaluation measures. The ranking experiment investigates the correlation among result lists that were calculated with the different semantic relatedness measures for a number of annotation systems. The stability experiment analyses the influence of noise in the ground truth on the ranked result lists. For this experiment, the binary ground truth annotations were randomly flipped from zero to one and vice versa for 1%, 2%, 5%, and 10% of the set. The evaluation score is calculated by using the altered ground truths and the correlation of the rankings is analysed. The overall goal is to determine which semantic relatedness measure displays the best characteristics for multi-label evaluation. This setup follows related work on evaluating performance measures (see Chapter 4.6), but introduces a new approach to measure stability.

8.3.1. Evaluation framework

The experiments are conducted in an evaluation framework that is schematically illustrated in Figure 8.1. The framework was already introduced in Chapter 7 and implements the OS evaluation measure. The core of the evaluation framework is the *matching procedure* that matches labels of the predicted set \mathcal{Z} to the ground truth set \mathcal{Y} and vice versa, according to Equation 7.1. The matching procedure takes a *costmap*, an *ontology*, and an *agreement map* into account as pluggable information resources. These information resources are further denoted as *plug-ins*. The predicted labelset \mathcal{Z} and the ground truth labelset \mathcal{Y} serve as input for each image and the framework outputs a score that describes the annotation quality by applying Equation 7.3.

The costmap plug-in is the most important part of the evaluation framework for this study. It describes the costs between pairs of concepts in the case of misclassification and can be determined in various ways. Originally, the OS used as costmap the DDMC, as explained in Chapter 7.3. As the experiments deal with a classification task, the vocabulary of concepts is fixed from the beginning and consists of 53 concepts in this study. The vocabulary is used to build a costmap for each pair of concepts. The costmap is represented as a confusion matrix and is symmetric. The costs are defined in the range of $[0, 1]$, where 1 determines the highest cost and 0 indicates no cost or equality of concepts. In the experiments, 14 semantic relatedness measures are investigated that were introduced in Section 8.2 (www_G, HSO, JCN, LCH, LESK, LIN, PATH, RES, VEC, WUP, WIKI, www_Y, FCS, and FTS) and turned into a costmap. The semantic relatedness measures were normalised and the relatedness value is converted into a cost by subtracting it from 1. These semantic costmaps are compared to the original proposed costmap of the OS, which realises the cost function of Equation 7.2, the annotator agreements, and the ontology knowledge. The measure is called HS if the ontology and the annotator agreement factors are not used.

¹This number refers to the amount of photos uploaded on Flickr in November 2009 when the experiment was conducted.

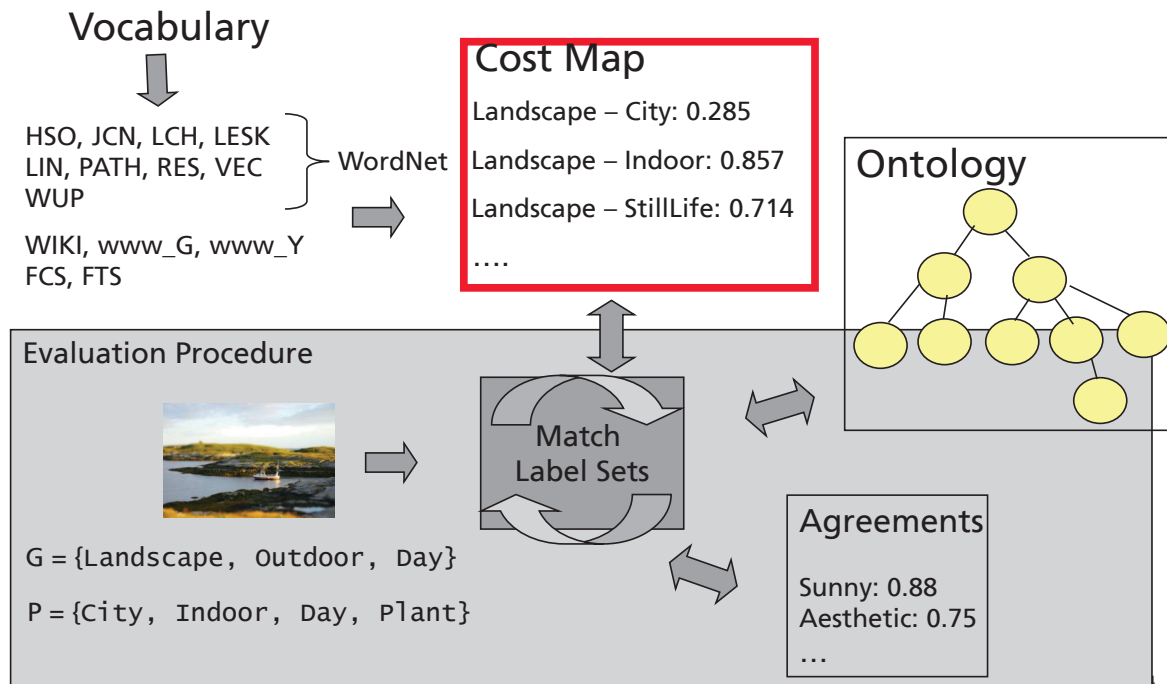


Figure 8.1.: Schematic representation of the evaluation framework.

8.3.2. Data

The experiments are carried out on the results of the runs of the ImageCLEF 2009 Photo Annotation task (see Chapter 10). In this task, 13,000 Flickr photos were annotated with 53 visual concepts by 19 research teams in 73 run configurations and one random run. The visual concepts were part of the PTO, as defined in Chapter 5. Each run is an unordered list containing the IDs of 13,000 test images followed by the confidence score, which gives an indication of the probability of each concept being present in an image. Initially, the confidence score is a floating point number between zero and one, where higher numbers denote higher confidence. In agreement with the participants, the confidence values were mapped to binary values using a threshold of 0.5 for the evaluation measures that need a binary decision about the presence of concepts. The utilisation of the ImageCLEF runs allows for a comparison of the semantic relatedness measures in a realistic annotation scenario, and offers diverse and numerous configurations and systems. In the experiments, the 15 introduced costmaps including the OS are plugged into the evaluation framework as described in the previous section. The scores for the 13,000 test images per run are averaged and ordered in a ranked list for each costmap. Not all concepts of the vocabulary of the Photo Annotation task are present in WordNet. To cope with this problem, some concepts were replaced by similar ones; for instance, Party instead of Partylife.

8.3.3. Configurations

In the experiments, two configurations of the evaluation measure are investigated. In the first configuration, each costmap is included in the evaluation procedure together with the ontology plug-in and the agreement plug-in. This configuration is further denoted as *complete measure*, as all parts of the evaluation framework are used. The second configuration explores the characteristics of each costmap without the other plug-ins. This means that the matching procedure solely considers the costmaps to assign costs for misclassifications. In the following, this configuration

is referred to as *costmap measure*. As baseline for comparison, the example-based F-measure (F) is used to rank the results. It showed convincing characteristics in example-based multi-label evaluation, as its score is not majorly influenced by random annotations or by the number of labels annotated per image (see Chapter 7.5). Further, it also belongs to the example-based measures that consider binary system predictions, such as the OS with semantic relatedness costmaps.

8.4. Study results

In the ranking experiment, the correlation between pairs of rankings of the ImageCLEF runs is analysed. For each relatedness measure, the ImageCLEF runs are evaluated and ordered into a ranked list, before the correlation is calculated between each pair of lists. The Kendall τ correlation coefficient as well as the Kolmogorov-Smirnov D statistic are employed in order to support the decisions of one statistic (see Chapter 4.6.1).

The second experiment analyses the stability of the different relatedness measures concerning noise in the ground truth. After evaluating the ImageCLEF runs, the rank correlation is investigated for each result list in comparison to the ranking with correct ground truth, and in comparison to the ranking at the previous stage of noise.

8.4.1. Ranking results

Table 8.1 shows the results for the ranking experiment. In the upper triangle, the correlations for the complete measures are depicted, while the lower triangle presents the Kendall τ coefficient for pairs of rankings of the costmap measures. The last row and the last column show the correlations to F. The cells that are coloured in grey demonstrate the pairs of measures for which the Kolmogorov-Smirnov test supported the Kendall τ decision for concordance. For the rankings of the complete measures, the coefficient is very high, with an average correlation for all pairs of 0.92. For all costmap measures, the correlation to other costmap measures is lower, with an average of 0.86. In contrast, the Kolmogorov-Smirnov test supports the decision in just 39% for the complete measures and in 21% for the costmap measures. The Kolmogorov-Smirnov test decides for concordance with the ranking of the F-measure in 6 of 15 cases for the complete measures, although the correlation coefficient of the Kendall test is low. In the case of the costmap measures, a correlation is supported in 3 of 15 cases. The ranked result lists change more significantly when applying different costmap measures.

Figure 8.2 visualises the Kendall τ correlation coefficients in a dendrogram after applying a binary hierarchical clustering. A dendrogram is a tree visualisation in which each step of the hierarchical clustering is represented as a fusion of two branches into a single branch. The dendrogram shows a similar clustering for both configurations. In both cases, the highest correlation between any two measures can be found between RES and LIN. This leads to the conclusion that the differences between these measures are rather small, and that the different way of scaling the information content between terms leads to an almost equal ranking. In the case of the costmap measures, HS has the lowest correlation to all other measures and falls into the outer cluster of the tree. For the complete measures, the F-measure shows the lowest correlation. Considering that the F-measure does not take into account ontology, agreements, and the matching procedure, which assigns fine-grained costs, this is not surprising. Interestingly, the F-measure behaves very similar to LESK in case of the costmap measures. The thesaurus-based measures behave quite similar, only LESK has a greater distance to the others and is clustered near the distributional methods. Also, WIKI as the thesaurus-based measure with a different corpus stays close. For the distributional methods, FCS and FTS behave in almost the same manner. In this experiment, the point of a better word coverage of FCS does not influence the results to a great

Table 8.1.: Kendall τ correlation coefficient between ranking of runs evaluated with the different semantic relatedness evaluation measures. The upper triangle shows the results for the complete measures and the lower triangle depicts the results for the costmap measures only. As baseline for comparison, the example-based F-measure (F) is illustrated in light grey. The cells coloured in grey illustrate the combinations where the Kolmogorov-Smirnov test showed concordance in the rankings.

	www_G	HSO	JCN	LCH	LESK	LIN	HS/OS	PATH	RES	VEC	WIKI	WUP	www_Y	FCS	FTS	F
www_G		.908	.859	.925	.859	.931	.860	.879	.941	.904	.906	.943	.933	.910	.893	.615
HSO	.855		.946	.965	.925	.952	.826	.964	.953	.973	.958	.898	.935	.931	.918	.674
JCN	.813	.910		.917	.932	.904	.775	.976	.902	.929	.930	.847	.893	.893	.889	.690
LCH	.889	.956	.890		.901	.959	.831	.941	.959	.963	.966	.927	.958	.932	.919	.648
LESK	.933	.869	.832	.886		.899	.803	.922	.899	.913	.914	.844	.900	.922	.930	.746
LIN	.882	.906	.836	.935	.877		.858	.923	.987	.963	.954	.941	.944	.935	.916	.645
HS/OS	.662	.610	.539	.629	.648	.685		.796	.859	.841	.831	.887	.853	.864	.850	.599
PATH	.835	.937	.968	.922	.850	.865	.565		.922	.950	.950	.870	.915	.911	.902	.675
RES	.885	.903	.835	.930	.881	.975	.673	.859		.956	.952	.944	.945	.942	.922	.644
VEC	.884	.930	.866	.957	.880	.936	.651	.891	.921		.959	.908	.931	.927	.917	.661
WIKI	.876	.882	.859	.913	.873	.893	.639	.877	.876	.905		.910	.940	.930	.920	.664
WUP	.856	.828	.757	.864	.827	.913	.721	.787	.913	.870	.824		.922	.904	.887	.599
www_Y	.902	.868	.809	.881	.910	.873	.699	.835	.870	.881	.859	.833		.934	.922	.650
FCS	.929	.839	.779	.860	.934	.876	.698	.801	.885	.861	.847	.859	.916		.978	.688
FTS	.916	.819	.767	.839	.922	.850	.708	.786	.857	.845	.836	.833	.910	.969		.706
F	.921	.867	.847	.883	.975	.870	.629	.859	.874	.873	.861	.817	.899	.918	.906	

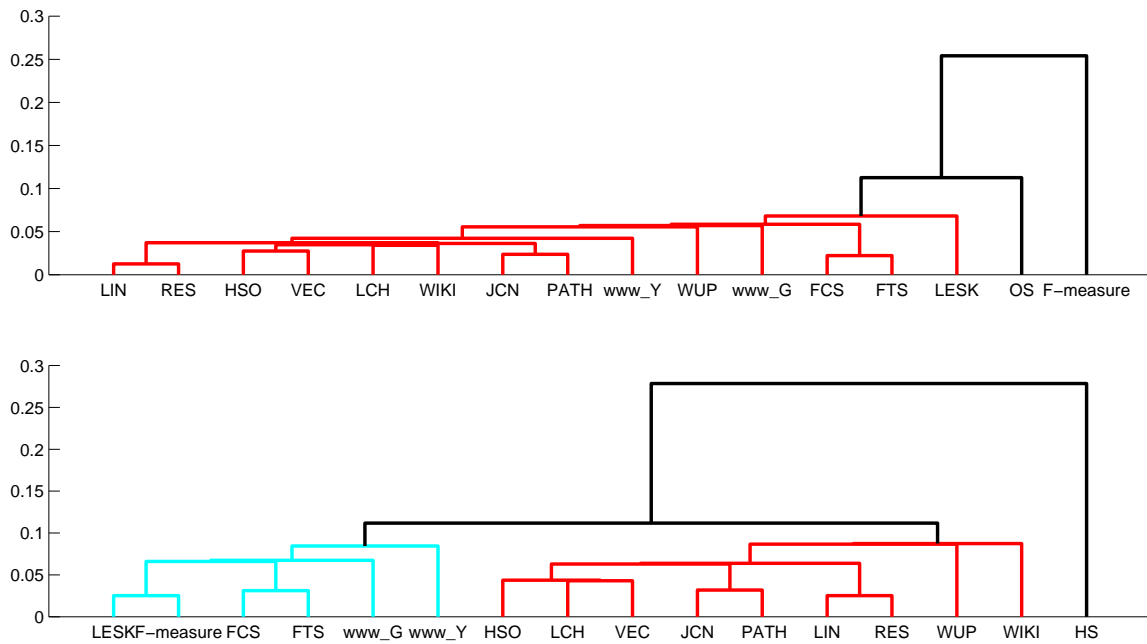


Figure 8.2.: The upper dendrogram shows the results after hierarchical classification for the complete measures, the lower one for the costmap measures.

extent. Summarising, the dendrogram shows that the plug-ins of the framework, while affecting the ranking more in case of the costmap measures, maintain the characteristics of the measures to each other with the exception of the F-measure.

8.4.2. Results of stability experiment

Table 8.2 illustrates the results of the stability experiment for the complete measures and the costmap measures. The table shows the ranking correlation for the complete measures to the original ranking, the costmap measures to the original ranking, the complete measures to the previous stage of introduced noise, and the costmap measures compared to the previous stage of noise from left to right. Again, the numbers in each cell denote the Kendall τ correlation coefficient and the grey cells highlight the combinations in which the Kolmogorov-Smirnov test supported the Kendall τ decision for concordance. In the last row, the results for the F-Measure are presented. The F-Measure is not computed using the evaluation framework and matching procedures. Therefore, there are no different results for costmap or complete measures.

For all measures, the correlation coefficient decreases with increasing amount of noise. As can be seen from the results of the ranking experiment, the Kendall τ coefficient is not very sensitive. Again, it supports a decision on correlation for every pair of rankings, although the correlation coefficient decreases to 0.38 at minimum. In the results of the Kolmogorov-Smirnov test, it is obvious that `www_G` changes the order of systems significantly by introducing only 1% noise for the complete measures compared to the original ranking. The OS measure shows a good stability, as it keeps a concordant ranking until 10% of noise are introduced. All other complete measures remain stable in ranking until more than 2% of noise are included in the ground truth. When the ranking of the complete measures is compared to the previous stage of noise, the OS and WUP remain stable over the four stages. `www_G` drops to discordance at the first stage, but is then concordant between the first and the second stage. The Kendall τ correlation coefficient is very high in this scenario, with over 0.9 correlation between the different stages. The costmap

Table 8.2.: The table depicts the Kendall τ correlations for the complete and the costmap measures between the original ranking and the ranking with altered ground truth on the left. A certain percentage of noise (1%, 2%, 5%, and 10%) was introduced into the ground truth. On the right, the correlations are shown when compared between the rankings of two noise stages. The cells coloured in grey illustrate the combinations where the Kolmogorov-Smirnov test showed concordance in the rankings.

	complete to original					costmap to original					complete to previous					costmap to previous				
	1%	2%	5%	10%		1%	2%	5%	10%		1%	2%	5%	10%		1%	2%	5%	10%	
www_G	.947	.954	.958	.931		.755	.746	.698	.604		.947	.989	.974	.954		.755	.973	.914	.852	
HSO	.990	.984	.953	.896		.722	.713	.687	.590		.990	.993	.968	.943		.722	.976	.927	.840	
JCN	.996	.986	.956	.927		.689	.682	.671	.624		.996	.990	.970	.971		.689	.980	.942	.907	
LCH	.989	.980	.938	.882		.730	.730	.673	.528		.989	.991	.958	.944		.730	.975	.891	.790	
LESK	.991	.984	.955	.893		.730	.721	.667	.567		.991	.992	.970	.938		.730	.979	.900	.839	
LIN	.983	.965	.919	.828		.739	.718	.641	.452		.983	.982	.954	.909		.739	.959	.859	.738	
OS/HS	.986	.968	.910	.829		.712	.682	.582	.379		.986	.982	.941	.919		.712	.942	.826	.753	
PATH	.993	.978	.956	.906		.698	.698	.675	.571		.993	.985	.978	.950		.698	.975	.939	.834	
RES	.981	.961	.910	.831		.749	.726	.638	.480		.981	.981	.949	.921		.749	.958	.856	.763	
VEC	.987	.975	.930	.853		.712	.704	.629	.478		.987	.988	.956	.923		.712	.968	.869	.761	
WIKI	.987	.973	.927	.836		.728	.709	.653	.438		.987	.986	.953	.910		.728	.962	.876	.705	
WUP	.985	.964	.910	.834		.756	.721	.624	.412		.985	.979	.945	.924		.756	.946	.841	.743	
www_Y	.992	.981	.958	.912		.753	.747	.708	.610		.992	.989	.977	.954		.753	.976	.925	.855	
FCS	.989	.974	.944	.893		.959	.933	.826	.644		.989	.985	.970	.949		.959	.974	.893	.819	
FTS	.992	.974	.942	.876		.959	.921	.811	.621		.992	.982	.967	.935		.959	.962	.890	.810	
F	.978	.954	.893	.773		-	-	-	-		.978	.976	.939	.879		-	-	-	-	

measures all behave equally when compared to the original ranking. The Kolmogorov-Smirnov test assigns concordance as long as not more than 2% noise are incorporated into the ground truth. At the same time, the Kendall test shows that the correlation coefficient varies significantly between the different measures at the stage of 10% noise. In case the costmap measures are compared to the previous stage of noise, the HS is again concordant. All other measures are stable in their ranking until more than 2% noise are incorporated. It is obvious that the correlation coefficient drops for all measures in the stage of 1% compared to the original except for FCS and FTS, but then rises again in the comparison between the other stages. The F-Measure acts similarly to most of the other measures by tolerating 2% noise without a major influence on the ranking, but with a drop in correlation with greater amount of noise.

In the following, an example for the ranking of `www_G` and `www_Y` in the configuration of the complete measure after 1% noise was introduced is analysed. The Kendall test assigns a correlation of 0.947 and 0.992, respectively, but the Kolmogorov-Smirnov test only decides for concordance in case of `www_Y`. Having a look at the first 20 ranks of the system ranking after introducing 1% noise, for `www_G` the order changes to (1, 2, 3, 7, 9, 4, 5, 12, 6, 10, 14, 8, 11, 15, 13, 17, 18, 16, 19, 32). In contrast, the order of the first 20 ranks of `www_Y` is permuted to (1, 2, 3, 4, 5, 8, 6, 7, 9, 10, 11, 12, 15, 13, 16, 14, 17, 19, 18, 20). One can see from these sequences of numbers that in the case of `www_G` the numbers are exchanged to a greater extent and swapped with more distant ranks than in the case of `www_Y`.

8.4.3. Discussion

To summarise, the ranking experiment exhibits a correlation for the thesaurus-based measures HSO, JCN, PATH, and VEC, and the image-based distributional measures FCS and FTS in comparison to the baseline measure for the complete measures. In the case of the costmap measures, the correlation only could be assigned to HSO, JCN, and PATH. The evaluation framework with all plug-ins therefore seems to push the relatedness measures closer to the baseline, as the ontology plug-in incorporates penalties in the case of violations. These penalties assign the maximum costs; the same values that the F-Measure assigns to incorrect classified labels. Regarding the stability experiment, the above mentioned measures performed rather well and at least 2% noise could be incorporated without significantly changing the order of systems. The distributional document-based method `www_G` could not convince in its results, as it reacts in an unstable manner to a small amount of noise and has no confirmed correlation of both tests in the ranking experiment to the baseline or related measures. The OS showed the longest stability and therefore tends to be not sensitive enough for changes in annotations.

8.5. Flickr as source for semantic relatedness

Despite the ranking correlations and stability characteristics in image annotation evaluation, it is essential that the semantic relatedness metrics cover the semantic relatedness of concepts as closely as possible to human opinion. Otherwise, the usage of fine-grained misclassification costs looses its grounding and only introduces noise into the evaluation process. Usually, the effectiveness of semantic relatedness measures is assessed based on human correlation experiments. Several studies have been conducted, for example in Rubenstein and Goodenough (1965), Miller and Charles (1991), Finkelstein et al. (2001), and Zesch and Gurevych (2010). However, these studies rely on the notion of semantic similarity rather than focusing on semantic relatedness. As a result, Gracia and Mena (2008) have collected a dataset with human judgements on semantic relatedness between pairs of words. In the following, an experiment is performed which determines to what extent different semantic relatedness measures correlate to human judgements. For the

WordNet measures and the distributional measures, these experiments have been performed in related work. However, it has not been investigated how the Flickr resource behaves in comparison to human judgements for semantic relatedness. This is especially important because the stability and ranking experiments have shown the superior performance of FCS and FTS in image annotation evaluation. Therefore, the correlations of human judgements to semantic relatedness measures are compared to those of the FTS and FCS measure.

8.5.1. Correlation of human judgements to semantic relatedness measures

A widely used test set on semantic similarity was proposed by Miller and Charles (1991). They defined 30 word pairs and asked 38 human judges to assess the similarity degree on a scale from 0 (no similarity) to 4 (perfect synonymy). This test tackles the notion of semantic similarity which includes synonyms and hypernyms. Relations such as meronyms, antonyms, or frequent associations are not covered, as it would be necessary in the case for semantic relatedness measures. Therefore, Gracia and Mena (2008) designed a similar test which focuses on semantic relatedness rather than semantic similarity and includes these types of relationships. They asked 30 humans to judge the semantic relatedness of 30 word pairs on the same scale from 0 to 4. The word pairs and averaged human correlations are illustrated in Table 8.3.

Gracia and Mena (2008) use the Spearman rank correlation coefficient to compare WordNet based measures with distributional measures on their test set. They favour the Spearman correlation coefficient over the Pearson correlation coefficient, as some measures, such as LESK, produce highly non-linear results. The results are illustrated in Table 8.4. In summary, they found that most Web-based approaches outperform the traditional WordNet measures, although there are two search engines for which the performance is rather low. Supporting to our results of the ranking and stability experiment, the semantic relatedness determined with the Yahoo search engine shows superior performance than that obtained with Google.

I completed the experiment by integrating the FCS and FTS measures, which are based on a visual corpus, and extended the number of WordNet measures by including JCN and PATH. Both WordNet measures are missing in the original work of Gracia and Mena (2008), but showed good performances in the ranking and stability experiments. The results are highlighted in bold in Table 8.4. JCN and PATH both show a very low correlation to human judgements. They do not seem to capture the characteristics of semantic relatedness well. On the other hand, both Flickr measures perform well and act superior to the WordNet measures. The wider coverage of words between the FCS and FTS can be clearly seen in this experiment. The FCS measure also performs superior to semantic relatedness scores obtained with Yahoo and Google search engines.

8.5.2. Semantic relatedness of visual concepts

The previous experiment on the Gracia and Mena (2008) dataset has shown that the Flickr based semantic similarity measures represent adequate means of determining semantic relatedness. However, the word pairs of the dataset are very different from visual concepts in image annotation. Many words do not have a visual expression, such as “yes,” “no,” “hour,” “minute,” “theorem,” or “soul,” amongst others, and are therefore not used to annotate images.

The semantic relatedness measures find their application in the evaluation of visual annotation approaches in this thesis. In this context, the correlation of semantic relatedness measures to human judgements for visual concepts is crucial. As a result, a new dataset of visual concept pairs is constructed and judged by humans. Similarly to the experiment of Miller and Charles (1991) and Gracia and Mena (2008), 30 word pairs are collected. The words refer to visual concepts that are defined in the first version of the PTO. Much like the other experiments, one word pair

Table 8.3.: Human judgement on semantic relatedness of the Gracia and Mena (2008) dataset and the one on visual concepts collected in this study.

Gracia and Mena dataset			Relatedness of visual concepts		
Word1	Word2	Score	Word 1	Word 2	Score
<i>person</i>	<i>person</i>	4.00	<i>night</i>	<i>night</i>	4.00
hour	minute	3.38	water	river	3.73
mathematics	theorem	3.30	plant	flower	3.58
blood	transfusion	3.28	tree	plant	3.53
keyboard	computer	3.25	clouds	sky	3.50
citizen	city	3.24	outdoor	landscape	3.03
river	lake	3.19	city	building	3.03
letter	message	3.16	river	lake	2.95
car	driver	3.14	beach	sunny	2.93
car	wheel	3.02	sky	night	2.08
ten	twelve	3.01	indoor	outdoor	2.00
yes	no	3.00	day	sky	1.98
penguin	Antarctica	2.96	overexposed	underexposed	1.93
pencil	paper	2.90	animals	food	1.93
sea	salt	2.87	no person	still life	1.90
person	soul	2.84	city	vehicle	1.65
computer	calculator	2.81	several persons	portrait	1.53
atom	bomb	2.63	out of focus	high quality photo	1.40
dog	friend	2.51	portrait	mountains	1.08
professional	actor	2.12	family	one person	1.03
theft	house	1.99	canvas	sunset	1.03
city	river	1.85	desert	sea	0.85
power	healing	1.25	aesthetic	winter	0.80
pen	lamp	0.65	city	river	0.78
theorem	wife	0.34	flower	city	0.55
cloud	computer	0.32	night	snow	0.55
blood	keyboard	0.12	landscape	indoor	0.53
nanometer	feeling	0.11	macro image	animal	0.53
xenon	soul	0.07	sea	party	0.20
transfusion	guitar	0.05	sunset	family	0.10

Table 8.4.: Correlations to Gracia and Mena dataset. The scores are extracted from Gracia and Mena (2008), except the ones printed in bold.

Web-based Measure	Value	WordNet Measure	Value	Flickr Measure	Value
Exalead	0.78	VEC	0.62	FCS	0.76
Yahoo	0.74	RES	0.56	FTS	0.70
Altavista	0.74	LESK	0.56		
Ask	0.72	WUP	0.47		
Google	0.71	HSO	0.46		
Live Search	0.44	LIN	0.46		
Clusty	0.41	LCH	0.41		
		PATH	0.41		
		JCN	0.40		

What is the HIT about?

We would like to measure how much word pairs are semantically related to each other. The relation is free and can be of any kind as parent-child relations, conditions (one word implies the appearance of another), similar meanings, totally unrelated and so on.

Please select for each word pair a slider position that determines the degree of semantic relatedness.

- 0: no relation (e.g. transfusion - guitar)
- 1: low relation (e.g. power - healing)
- 2: moderate relation (e.g. theft - house)
- 3: high relation (e.g. car - driver)
- 4: strong or identical relation (e.g. car - automobile)

Imagine a scenario where these words refer to **instances in photos** or to a **description of photos**. How semantically close are these pairs?

<p>city - vehicle</p> <p>Please answer.</p>  <p>not related strongly related</p>	<p>portrait - mountains</p> <p>Please answer.</p>  <p>not related strongly related</p>	<p>water - river</p> <p>Please answer.</p>  <p>not related strongly related</p>
<p>outdoor - landscape</p> <p>Please answer.</p>  <p>not related strongly related</p>	<p>landscape - indoor</p> <p>Please answer.</p>  <p>not related strongly related</p>	<p>animals - food</p> <p>Please answer.</p>  <p>not related strongly related</p>
<p>day - sky</p> <p>Please answer.</p>  <p>not related strongly related</p>	<p>sunset - family</p> <p>Please answer.</p>  <p>not related strongly related</p>	<p>desert - sea</p> <p>Please answer.</p>  <p>not related strongly related</p>

Please provide any comments you may have below, we appreciate your input!

Figure 8.3.: Excerpt of the MTurk survey on visual concept correlations. For visualisation purposes just 9 out of 30 word pairs are depicted.

includes identical concepts, while others are specialisations, opposite terms, frequent associations, or determine part-of relations. The 30 word pairs are depicted in Table 8.3 on the right. Two word pairs of the set from Gracia and Mena (2008) are reused (river-lake and city-river). All words are judged on a scale from 0 (no relation) to 4 (strong or identical relation).

In this study, the word pairs are assessed according to their semantic relatedness by 40 persons. The assessment is performed as survey at MTurk. A total of 40 distinct turkers judged the complete dataset. All turkers that took part in the survey are located in the US. While there is no guarantee that every turker has English as mother tongue, there is a high probability that the turkers located in the US are able to speak English fluently and have an inherent understanding of the meaning of words. The design of the HIT assures that every word pair of the survey is judged by controlling the completion grade of the input before a submission is accepted. This approach is advantageous because it poses the same expenses for real answers as for spamming. The turkers therefore do not obtain a time advantage from cheating. All turkers that did not judge

Table 8.5.: Spearman correlations of semantic relatedness measures to the visual concepts dataset.

Web-based	Value	WordNet	Value	Flickr	Value	Other	Value
www_Y	0.633	WUP	0.507	FTS	0.799	OS	0.713
www_G	0.405	JCN	0.451	FCS	0.752	WIKI	0.396
		PATH	0.449				
		LIN	0.424				
		HSO	0.405				
		RES	0.357				
		LCH	0.310				
		VEC	0.232				
		LESK	0.060				

the identical word pair “night-night” with 4 were rejected, and these surveys are not considered in the analysis. Figure 8.3 shows part of the MTurk survey. The word pairs in each survey are arranged randomly to eliminate sequence effects. The average semantic relatedness of the word pairs is included in Table 8.3.

Table 8.5 shows the Spearman correlations of different semantic relatedness measures to the human judgements on the visual concept dataset. The Flickr measures inhibit the highest semantic relatedness to human opinion on this dataset. Interestingly, FTS performs superior to FCS. The tags seem to have characteristics which are closer to the visual concepts than to the whole Flickr context. The wider coverage of words in FCS does not prove to be advantageous in the determination of semantic relatedness of visual concepts. On the contrary, it introduces noise. The cost function of the OS is the third closest measure to human semantic relatedness judgement. Although there are issues with the OS as illustrated in Section 8.1, the semantic relatedness is captured, on the whole, quite well in the PTO. One has to note that one obstacle of the PTO, which judges the distance between “landscape-indoor” in the same way as the distance between “landscape-outdoor,” is regarded differently by humans. While they account for a high relatedness between “landscape-outdoor,” the relatedness between “landscape-indoor” is judged as rather low.

The WordNet measures do not perform well on the visual concept dataset. The measures WUP, JCN, and PATH are closest to human judgements. These results are contrary to the semantic relatedness judgement on the Gracia and Mena (2008) dataset. JCN and PATH performed worst on their set. However, the results on the visual concept dataset are in accordance with the good characteristics of JCN and PATH in the stability and ranking experiments. Again, the distributional measure using the Yahoo search engine performs superior to that using Google. In contrast to the results of Gracia and Mena (2008), the Web-based measures do not outperform the WordNet measures as clearly on the visual concept dataset. The measure WIKI does not show a high correlation to human judgements.

8.6. Summary

In this chapter, the behaviour of several semantic relatedness measures for the evaluation of multi-label image classification was studied. The 15 semantic relatedness measures are based on WordNet, Wikipedia, Flickr, and on the WWW, and were included in the OS as plug-ins to determine the costs for misclassification. Their performance was compared to the example-based F-Measure in two experiments, the ranking and the stability experiment. In summary, the thesaurus-based measures HSO, JCN, and PATH, and the image-based distributional measures FCS and FTS account for the highest correlation in the ranking experiment, especially in the

complete configuration. Summarising the stability experiment, the measures are stable in their ranking for nearly all configurations until more than 2% of noise are introduced. `www_G` acts in an unstable manner from the beginning. The original OS shows a longer stability in the ranking than the other measures. It has to be investigated whether it is sensitive enough in its ranking to cope with noise. Even though the mentioned WordNet based measures showed promising results, the vocabulary had to be adapted as not all words of the vocabulary were present in WordNet.

The second part of the chapter deals with the correlation of semantic relatedness measures with human judgement on semantic relatedness of word pairs. The measures were evaluated on established datasets, and in particular, the Flickr based measures performed in a superior fashion. Due to the lack of a dataset on visual concepts, a new dataset was collected and assessed by 40 turkers in a MTurk experiment. Again, the Flickr based measures performed in a superior fashion and, in particular, the FTS accounts for a high correlation. Therefore, the usage of the FTS measure as cost function in the OS is recommended for the evaluation of visual annotation approaches.

A limitation of this work is the comparison of the relatedness measure ranking characteristics that consider fine-grained costs between predicted and ground truth annotations with the F-Measure as baseline, that utilises binary scores. Naturally, these measures cannot and should not behave the same. The behaviour of measures should better be compared with human preferences on evaluation measures. While the determination of human judgement on the semantic relatedness of terms and the computation of correlations among relatedness metrics with human opinion determines a promising first step, the ranking and stability preferences of users should be taken into account. Such a large-scale user study is out of scope of this thesis which left me to choose the F-Measure as stable baseline in balance between Precision and Recall.

9. Exploiting semantics in a multi-label performance measure for ranked predictions



This chapter proposes a novel performance measure for ranked predictions that judges the quality of a set of annotated concepts. In the two preceding chapters, the OS measure for multi-label image annotation evaluation was discussed and extended in order to improve capturing semantic relatedness among visual concepts in accordance with human judgements. The resulting Ontology Score with Flickr Tag Similarity (OS-FTS) follows the proposed user model by incorporating the requirements on example-based evaluation and the usage of fine-grained evaluation costs. However, the OS-FTS requires the annotation systems to predict label sets with a clear decision on the presence and absence of concepts. The Semantic R-Precision (SR-Precision) tackles the other side of the prediction format dimension and is suited for ranked system predictions while addressing all user requirements. The effectiveness of the SR-Precision measure is experimentally evaluated on the runs submitted to the Photo Annotation task in 2009 and 2010. The chapter is structured as follows: Section 9.1 provides the motivation of the new performance measure by discussing difficulties in the application of set-based measures to ranked predictions. Section 9.2 reconsiders and extends the proposed user model for VIR and relates performance measures in use to its requirements. Section 9.3 introduces the SR-Precision measure, while Section 9.4 describes the experimental setup and the study results. The user model is verified in Section 9.5 with human judgements on image predictions, and the chapter closes with a short summary.

9.1. Motivation

Performance evaluation of annotation approaches usually adheres to established IR measures that evaluate each assigned concept in isolation. This setup ignores that different semantic concepts are indexed for the same multimedia document and that the answer to a complex query is given

by the set of indexed concepts. This issue has been extensively discussed in the previous chapters. While a few performance measures have been adapted for the example-based evaluation, evaluation measures for indexation based on ranked predictions are rarely in use, as shown in Chapter 4.3. Usually, concept-based Precision, Recall, MAP or inferred variants of MAP are applied.

The predictions retrieved in the annotation step of a retrieval system are binary decisions or likelihoods that determine how certain a system is about the annotation of a particular concept. The likelihoods can be converted into a ranking of concepts (possibly with ties) per media item. In the analysis of evaluation measure dimensions, we have seen that performance measures can be differentiated into label set measures and measures relying on ranked predictions. In the first case, the annotation system partitions the set of labels into relevant and irrelevant concepts. In the latter case, the cut-off value is left to the performance measure. Label set measures can be applied to the evaluation of annotation systems that generate likelihoods after a threshold maps the confidence values to a binary assignment. However, determining this threshold is difficult and it has to be chosen carefully. The thresholding can further obscure the performance of the classifier, as practically two processing steps are evaluated at once. Binary predictions are disadvantaged in the evaluation with rank-based measures, as the binary assignment cannot be converted into a proper ranking (see Chapter 7.5). By forcing the system to produce binary decisions, it further loses confidence information that might be useful in a later query processing step.

In the following, the thresholding problem is illustrated based on the submissions to the Image-CLEF Photo Annotation task in 2010. The participants were asked to submit binary predictions as well as likelihoods for a test set of 10,000 images and 93 concepts. This experiment considers two example-based performance measures: the F-measure (Equation 3.16) to evaluate the binary submissions and the R-Precision measure (Equation 3.25) to evaluate the ranked predictions. Both measures are based on Precision and Recall and have similar intentions. The runs are ranked according to the average score of the F-measure and the R-Precision measure, respectively.

Figure 9.1 presents the differences in the ranking of the runs. Overall, the two rankings correlate with a Kendall τ correlation coefficient of 0.77. While the first four ranks are only slightly swapped, extreme differences can be found for the ranking of other runs. One has to note that these first four runs correspond to submissions from the same research group that explicitly focused on the thresholding problem (van de Sande and Gevers (2010)). While some differences may result from the different perspectives of the evaluation measures, some changes probably result from an awkward threshold.

9.2. The need for yet another evaluation measure

In the following, the proposed user model is reconsidered and related to existing performance measures. As shown in the previous section, it might be advantageous to evaluate annotation systems directly based on the ranked predictions. Thresholding predictions adequately poses a separate research problem, so that this issue comprises an additional source of error. Further, the user might be satisfied with a result list containing a few dozen images, depending on the search objective. These images should be ranked according to their relevance to the query. Usually, the search component relies on the likelihood information of the classifier for ranking. By forcing the classifier to produce binary decisions for annotation evaluation, the system loses the confidence information it may need. Search components that still retain likelihoods for the ranking are evaluated on a different basis than those the search component uses in case a label set performance measure is applied. This leads to a third user requirement:

- *The measure should consider the predictions in the format that is used in further processing units of the search system.*

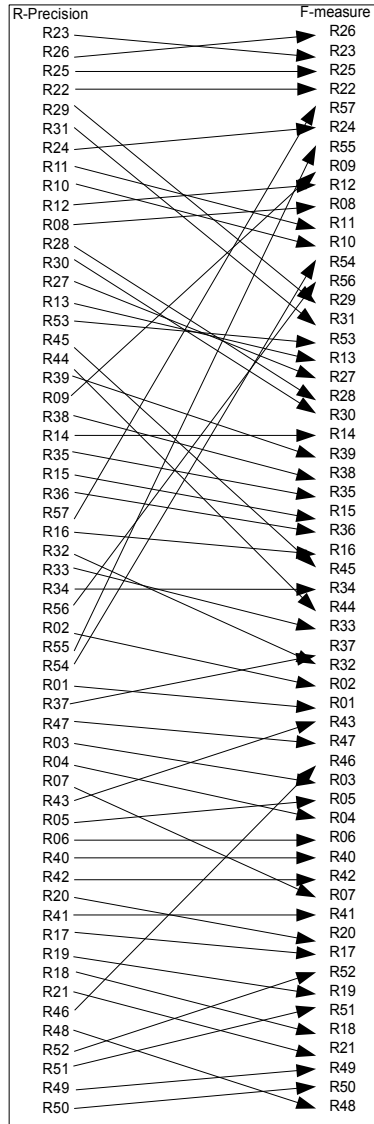


Figure 9.1.: Comparison of the ranking of runs in terms of example-based R-Precision and F-measure based on the confidence values and binary predictions, respectively.

In addition to the first requirement on assessing performance with measures that consider the complete set of ground truth labels, and the second requirement on the judgement of misclassifications in a fine-grained fashion, this requirement completes the user model for concept-based image search.

The example-based performance measures for label set predictions and ranked assignments introduced in Chapters 3.3 are discussed concerning their ability to satisfy the user model in the following. In particular, the measures One Error, Coverage, Ranking Loss, example-based Mean Average Precision (MAP-Ex), and R-Precision are considered for ranked assignments, while the measures Hamming Loss, Precision, Recall, and F-measure are adopted for label set predictions.

Starting with the first requirement, the measure One Error fails to fulfil the need to consider the whole ground truth set of labels. It considers only the label with the highest rank. In visual indexing, there are a few concepts that are practically depicted in a high percentage of images, such as the concepts *Day* or *Outdoor*. A classifier may predict these kinds of concepts with a high confidence, and One Error assigns high scores ignoring all other concepts and requirements

that were defined. The measure MAP-Ex shows a similar characteristic, although it considers the complete ground truth label set. Moffat and Zobel (2008) usefully demonstrate how much the document with highest rank influences the concept-based MAP. In their example of 20 retrieved documents, the influence of the first document contributes 58% to the overall score for a prediction of the list $\$ - - - \$ - - - - \$ - - - - - \$ - - -$, with $\$$ denoting a relevant document. Formulated for the MAP-Ex measure, this result is transferable to an annotation scenario of one image with 20 concepts. The question remains of why a concept from a set of concepts should influence the evaluation score in this way, recalling that all concepts are equally visible in the image.

All measures presented in Chapter 3.3.4 and 3.3.5 violate Requirement 2. The measures treat each misclassification equally and assign binary scores for the annotation performance. They all assume an independence of concepts in the concept sets. Finally, the label set measures (potentially) violate Requirement 3, as rankings have to be converted into binary decisions before the measures are applicable. However, this requirement clearly depends on the system setup.

Example-based multi-label classification evaluation measures that take the degree of misclassification into account were reviewed in Chapter 4.4. These measures make use of ontologies, hierarchies, or other external knowledge bases to provide an estimate of the degree of semantic relatedness between a predicted and a ground truth concept. These measures are not established in annotation evaluation, as shown in Chapter 4.3. Moreover, many need an extensive hierarchy or ontology to determine performance. In order to deal with these issues, I proposed the OS measure and its extension, the OS-FTS measure, in the Chapters 7 and 8, respectively. However, none of these measures are defined for ranked annotations, including those from related work. They all rely on binary predictions of the system. To sum up, there is no example-based performance measure that fulfils all requirements and considers fine-grained costs for misclassifications in use.

To account for these issues, a new performance measure, the SR-Precision, is proposed which is able to deal with ranked annotations. With respect to the results obtained in Chapter 8, the SR-Precision considers the FTS between concepts in order to differentiate between badly and moderately well annotated documents. The effectiveness of the SR-Precision measure is experimentally evaluated on the runs submitted to the ImageCLEF Photo Annotation task in 2009 and 2010, and compared to other example-based evaluation measures for ranked predictions.

9.3. Semantic R-Precision

In the following, I introduce the new evaluation measure. It is called SR-Precision, as it originates from the example-based R-Precision, but considers the semantic relatedness between concepts to deal with misclassifications. In contrast to the hierarchical and ontology-based measures as introduced in Chapter 4.4, it can be used in the absence of a hierarchy, taxonomy, or ontology.

In general, the definition of an example-based multi-label classification evaluation measure has to cope with the two issues of (1) how to determine the contributions for misclassifications to the overall score and (2) how to map the predicted label set to the ground truth label set. The computation of the contributions of misclassifications in the SR-Precision follows results from the previous chapter and uses the Flickr platform with the FTS as similarity function. The second issue of mapping the predicted label set to the ground truth label set is solved as follows:

SR-Precision considers the subset \mathcal{Z}_n^* of the predicted set \mathcal{Z}_n which contains as many elements as there are relevant concepts in the ground truth \mathcal{Y}_n :

$$\mathcal{Z}_n^* \subseteq \mathcal{Z}_n \Leftrightarrow \forall c \in \mathcal{Z}_n : \text{rank}(f(z_{nc})) \leq |\mathcal{Y}_n|. \quad (9.1)$$

The ranking function $\text{rank}(f(z_{nc}))$ sorts the concept predictions in \mathcal{Z}_n in descending order according to their likelihood value. So far, the measure SR-Precision equals the example-based

R-Precision. In contrast to R-Precision, the precision of the prediction is not calculated in a binary fashion, but the semantic relatedness between the concept sets is considered. Therefore, each concept label of \mathcal{Z}_n^* is assigned to one concept label of the ground truth \mathcal{Y}_n by maximizing the overall score. This problem is also known as *linear assignment problem*. The contribution each pairing gives to the overall score is determined by the FTS semantic relatedness measure. A high co-occurrence between the concepts denotes a high contribution to the overall score, while a low co-occurrence accounts for a low similarity and contributes less to the overall score. Consequently, the contribution of equal concept labels in both sets is 1. The qualification matrix q for an image d is constructed from all entries c_j of the reduced prediction \mathcal{Z}_n^* and c_k from the ground truth \mathcal{Y}_n :

$$q_{jk} = \text{FTS}(c_j, c_k). \quad (9.2)$$

The matrix q_{jk} is then used to determine the best assignments with respect to the highest semantic relatedness between concepts c_j and c_k . This assignment function $a_n(x)$ can be formulated as a maximization problem:

$$\begin{aligned} a_n(x) &= \sum_{j=1}^c \sum_{k=1}^c q_{jk} x_{jk} \longrightarrow \max, \\ \sum_{j=1}^c x_{jk} &= 1 \quad \forall k \in \{1, \dots, c\}, \\ \sum_{k=1}^c x_{jk} &= 1 \quad \forall j \in \{1, \dots, c\}, \\ x_{jk} &\geq 0 \quad \forall j, k. \end{aligned} \quad (9.3)$$

x_{jk} determines the assignment from the j^{th} element of set \mathcal{Z}_n^* to the k^{th} element of \mathcal{Y}_n . The linear assignment problem can efficiently be solved with the Hungarian method, also known as Kuhn-Munkres algorithm (see Kuhn (1955)). The final score for the SR-Precision is:

$$\text{SR-Precision}(Z, Y) = \frac{1}{N} \sum_{n=1}^N \frac{a_n(x)}{|\mathcal{Y}_n|}. \quad (9.4)$$

9.4. Experimental work

The experiments make use of the ImageCLEF 2009 and 2010 submissions to the Photo Annotation task. In 2009, 13,000 images were annotated with 53 concepts, while in 2010, 10,000 images were annotated with 93 concepts. Although, in both cycles, the participants were asked to submit likelihoods on the presence of concepts, not all groups followed this request. All runs with binary concept assignments are excluded in the experiments. This leads to a consideration of a total number of 45 runs submitted by 12 participating groups in 2009, and 57 runs submitted by 16 groups in 2010. Further, for each cycle, a submission with pseudo random numbers in the interval of $[0; 1]$ is added. The characteristics of the SR-Precision are compared to five other example-based performance measures for ranked predictions, namely Coverage, One Error, MAP-Ex, R-Precision, and Ranking Loss in three experiments:

1. Correlation analysis

Three correlation experiments are performed. First, the predictions for each image in each run are evaluated with the corresponding performance measure, and the images are ranked according to the evaluation score in each run. Then, the correlation between pairs of

measures for each run is evaluated and averaged over the runs. This experiment therefore investigates the correlations in the ranking of images among different measures. Second, the correlation between the ranking of systems is analysed. For each measure, all ImageCLEF runs are evaluated and ordered into a ranked list by their average annotation score. The correlation is then calculated between each pair of lists of the overall system ranking. Third, a visual correlation between the results of two measures is presented. Therefore, scatter plots are analysed that show the overall scores on the tasks for different pairings of measures.

2. Stability experiment

The second experiment analyses the stability of the different example-based measures concerning noise in the ground truth. After evaluating the ImageCLEF runs, the rank correlation is investigated for each result list in comparison to the ranking with correct ground truth, and in comparison to the ranking at the previous stage of noise. The setup of the stability experiment is comparable to the stability experiment in Chapter 8.4.2.

3. Comparison of performance assessment on the basis of image scores

The last experiment compares the performance scores assigned by SR-Precision to the ones assigned by R-Precision on an image basis. Scatter plots demonstrate effectiveness score distributions from 12 runs of ImageCLEF 2009 and 13 runs of ImageCLEF 2010. Each run corresponds to a submission from another team.

9.4.1. Results of the correlation experiment

The correlations of image rankings between pairs of measures are presented in Table 9.1. The upper triangle corresponds to the submissions in 2009 and the lower triangle to the submissions in 2010. Please note that the measures Coverage, One Error, and Ranking Loss assign better quality with lower scores, while the measures MAP-Ex, R-Precision, and SR-Precision reach their optimal score with the value 1. Therefore, the correlations are negative when comparing measures from both groups. It is obvious that the correlation between most measures is rather low. Disregarding the algebraic sign, an average correlation of 0.38 and 0.40 is reached in both years. R-Precision and SR-Precision show the highest correlation with an average correlation coefficient of 0.86. For the other measures, Ranking Loss is most highly correlated with Coverage, and One Error has the highest correlation to MAP-Ex and vice versa. These results hold for the submission of both years. The lowest correlation is between R-Precision, SR-Precision, and Ranking Loss with the measure One Error, while MAP-Ex and One Error correlate least to the measure Coverage.

These findings change when the correlations for the final ranking of runs are considered. Table 9.2 presents the Kendall τ correlation coefficients for the correlation of the system rankings. Again, the upper triangle depicts the results for ImageCLEF 2009 and the lower triangle those for ImageCLEF 2010. These correlations are significantly higher than those assessed between the ranking of images per run. It seems as if the measures agree on how to rank the systems in direct comparison to each other, despite the differences in assessing the annotation performance for single images. Again, the rankings show the same average correlations for both tasks with 0.837 and 0.835 disregarding the algebraic sign. In both years, the measures SR-Precision and R-Precision rank the systems almost identical with an average correlation coefficient of 0.97. Further, One Error correlates best with MAP-Ex (\varnothing 0.87), but MAP-Ex correlates higher to R-Precision in 2009. Coverage correlates best with Ranking Loss (\varnothing 0.86), but Ranking Loss has a higher correlation to R-Precision in 2010. Although the correlations show small differences between the two years, the general trend in the image ranking correlations holds also for the system rank correlations. The lowest correlation in 2010 is clearly reported from all measures to Coverage, while Coverage itself shows the lowest correlation to One Error. The results for 2009 are not as

Table 9.1.: The table depicts the averaged Kendall τ correlation of the image rankings between each run evaluated with several performance measures. The upper triangle shows the correlations for the rankings of the Photo Annotation task 2009 while the lower triangle presents the correlations for the rankings in 2010. The numbers marked in bold show the highest correlation for any measure to the other measures.

	Coverage	MAP-Ex	One Error	R-Precision	R-Loss	SR-Precision
Coverage	1	-0.246	0.162	-0.302	0.579	-0.288
MAP-Ex	-0.226	1	-0.553	0.429	-0.352	0.409
One Error	0.149	-0.602	1	-0.225	0.202	-0.214
R-Precision	-0.302	0.388	-0.238	1	-0.494	0.859
R-Loss	0.560	-0.371	0.223	-0.558	1	-0.442
SR-Precision	-0.294	0.371	-0.230	0.864	-0.523	1

Table 9.2.: The table depicts the Kendall τ correlation between the overall rankings of the submissions evaluated with several performance measures. The upper triangle shows the correlations for the rankings of the Photo Annotation task 2009 while the lower triangle presents the correlations for the rankings in 2010. The numbers marked in bold show the highest correlation for any measure to the other measures.

	Coverage	MAP-Ex	One Error	R-Precision	R-Loss	SR-Precision
Coverage	1	-0.785	0.761	-0.814	0.911	-0.797
MAP-Ex	-0.684	1	-0.819	0.896	-0.863	0.888
One Error	0.698	-0.915	1	-0.755	0.778	-0.749
R-Precision	-0.717	0.907	-0.855	1	-0.898	0.958
R-Loss	0.806	-0.866	0.817	-0.906	1	-0.882
SR-Precision	-0.707	0.914	-0.850	0.988	-0.896	1

distinct. R-Precision, Ranking Loss, Coverage, and SR-Precision show the lowest correlations to One Error, while MAP-Ex has the lowest correlation to Coverage, and One Error correlates least with SR-Precision. Overall, the high correlation of SR-Precision with R-Precision is observed in the image and system correlation experiment with 0.84 and 0.97, respectively.

These findings are visualised in Figure 9.2. The Kendall τ correlation coefficients are presented in a dendrogram after applying a binary hierarchical clustering. A dendrogram is a tree visualisation in which each step of the hierarchical clustering is represented as a fusion of two branches into a single branch. The measures are abbreviated with *Rp* for R-Precision, *Srp* for SR-Precision, *Map* for MAP-Ex, *Rl* for Ranking Loss, *Cv* for Coverage, and *Oe* for One Error. It can be seen that the relation between the measures is the same for the image rankings in 2009 and 2010, while there are differences in the relation of measures for the overall ranking. Nonetheless, the high correlation of SR-Precision with R-Precision is visually confirmed for all four experiments. Especially, in the system ranking for 2010, the measures nearly rank the systems identically.

The scatter plots in Figure 9.3 visualise the averaged effectiveness scores of all 45 submitted runs from 2009 in black dots and the 57 submitted runs from 2010 in blue crosses. Further, the two random configurations and the ground truth scores are shown in red and green, respectively. Exemplarily, pairs of evaluation measures were chosen and plotted against each other. All scores were converted so that higher scores correspond to better quality. For all measures, the random

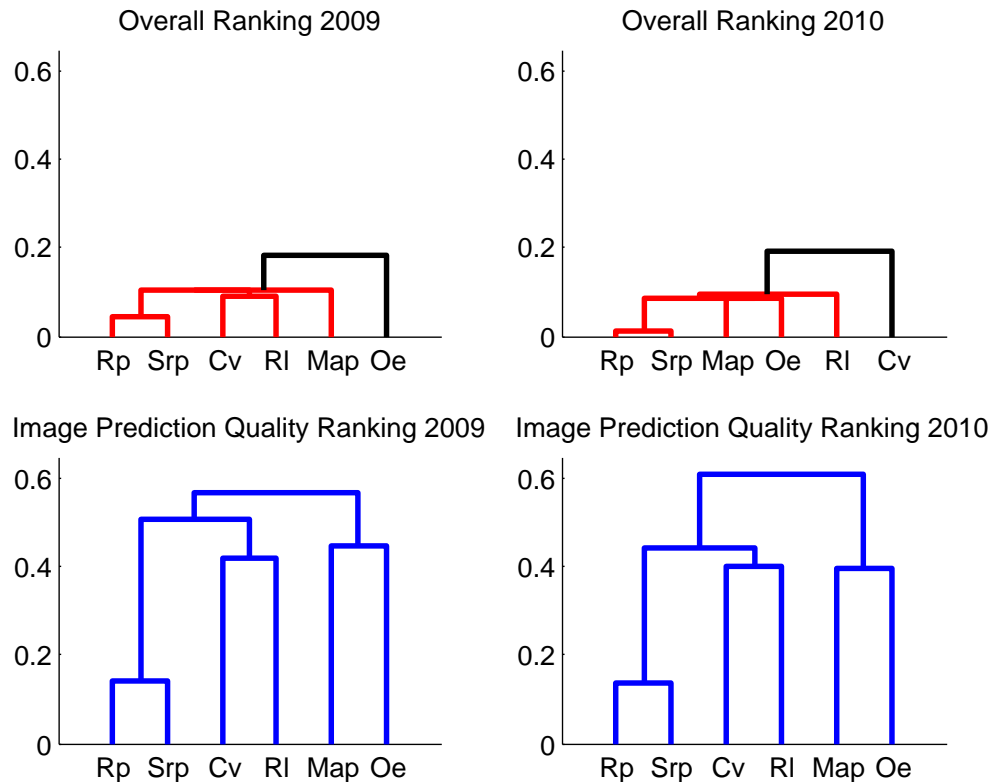


Figure 9.2.: The dendrograms visualise the relation between the performance measures based on the Kendall τ rank correlation coefficient. In the top left and top right plot, the results for the overall ranking in 2009 and 2010 are depicted, while the plots in the bottom show the results for the image rankings per run in 2009 and 2010.

configurations achieve low scores and are placed close to the end of the result list. This behaviour shows that all measures successfully differentiate between random runs and system performance in contrast to the concept-based F-measure as shown in Chapter 7.5. The first five plots a) - e) depict the scores for the SR-Precision in contrast to all other measures. The high correlation to R-Precision can clearly be seen in Subplot a). The results are correlated in a straight line. Further, results show that SR-Precision assigns higher scores than R-Precision, especially for systems with low performance. The notion of considering semantics for misclassifications relaxes the cost computation but still retains the ability to discriminate well among systems. Further, the similar behaviour of R-Precision and SR-Precision can be seen in the correlations to other measures, as shown in Subplot d) and j) in comparison to the measure MAP-Ex. While MAP-Ex is in general a quite strict measure regarding its score, the systems performing worst get higher scores with SR-Precision, but a stronger correlation can be seen for the top submissions. SR-Precision also shows a high visual correlation to Ranking Loss. While there is no difference in the scores of SR-Precision for 2009 and 2010, Ranking Loss clearly differentiates runs of both years in its score. The runs of 2010 get better scores assigned than the runs of 2009. The reason lies in the total amount of visual concepts in these years that increased from 53 concepts in 2009 to 93 concepts in 2010. At the same time, the average number of depicted concepts per image rose only from 9 to 12 concepts which means that the LD decreased from 0.17 to 0.13. Ranking Loss considers how often an irrelevant concept is ranked before a relevant one, divided by the total number of concepts. This score is more likely to result in a higher Ranking Loss for the 2009 task. Further, SR-Precision is almost uncorrelated to Coverage and to One Error. Subplot g) shows that Coverage and One Error are quite uncorrelated. This behaviour confirms

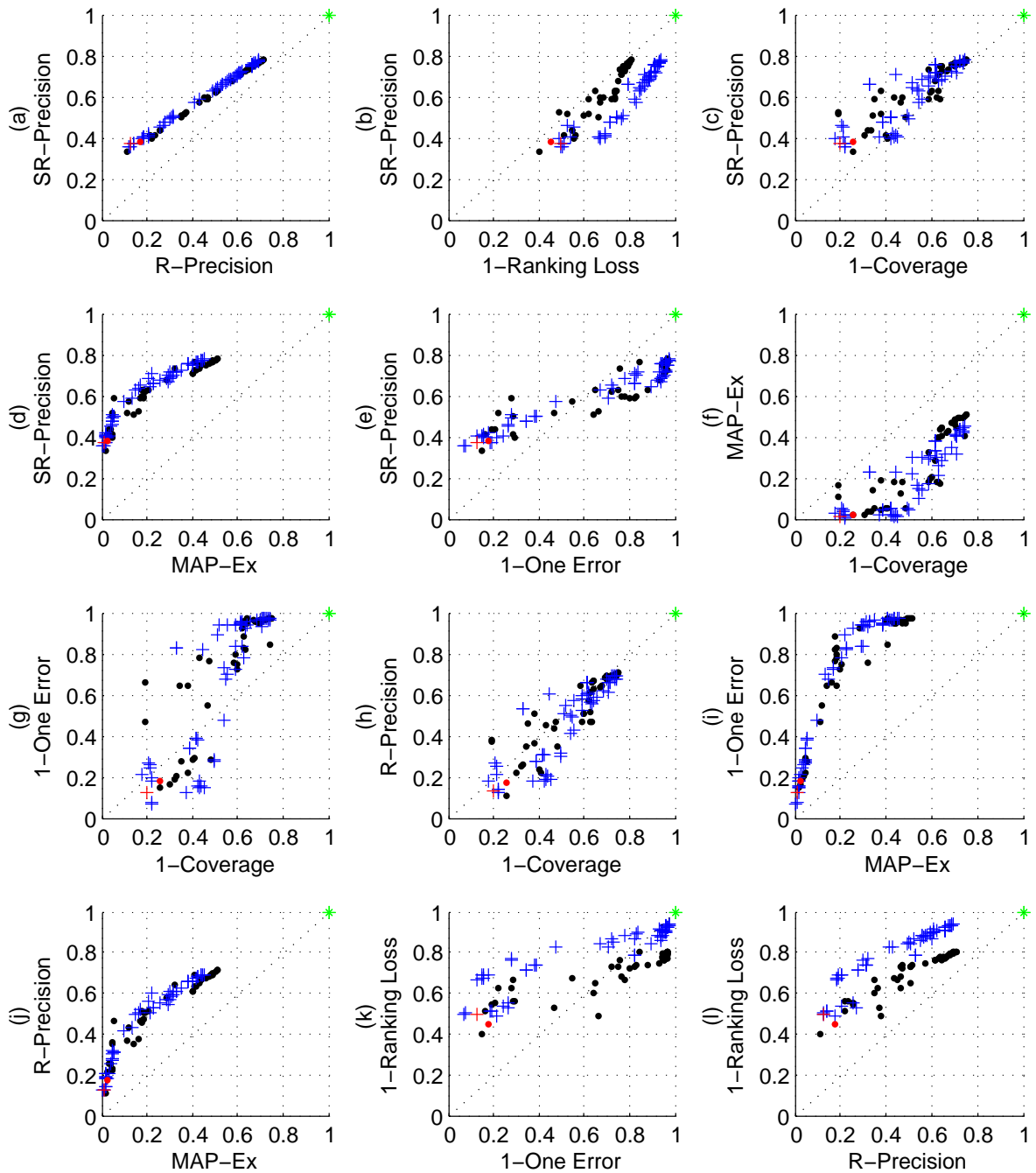


Figure 9.3.: The scatter plots visualise the results for the run configurations for pairs of evaluation measures. The black dots correspond to the runs of the Photo Annotation task in 2009, while the blue crosses denote the runs in 2010. One random configuration is visualised as a red dot and a red cross, respectively. The green stars show the best possible score for each measure. To make the visualisation intuitive, all measures are normalised so that higher scores denote better results.

Table 9.3.: The table depicts the Kendall τ correlations for the ranking of runs between the original ground truth and the altered ground truth by $x\%$ in the upper part. In the lower part, the correlations in the rankings between two noise stages are shown. Results for 2009 and 2010 are depicted on the left and right, respectively.

	to original 2009				to original 2010			
	1%	2%	5%	10%	1%	2%	5%	10%
Coverage	0.992	0.977	0.944	0.909	0.990	0.959	0.899	0.812
MAP-Ex	0.998	0.998	0.990	0.986	0.997	0.999	0.995	0.984
One Error	0.999	0.991	0.989	0.981	1	0.985	0.988	0.983
R-Precision	0.994	0.985	0.965	0.963	0.995	0.989	0.974	0.952
R-Loss	0.998	0.996	0.990	0.985	0.996	0.997	0.996	0.996
SR-Precision	0.985	0.977	0.961	0.913	0.997	0.990	0.979	0.971
	to previous 2009				to previous 2010			
	1%	2%	5%	10%	1%	2%	5%	10%
Coverage	0.992	0.985	0.963	0.965	0.990	0.968	0.923	0.859
MAP-Ex	0.998	0.996	0.992	0.992	0.997	0.998	0.994	0.986
One Error	0.999	0.991	0.989	0.981	1	0.985	0.985	0.979
R-Precision	0.994	0.986	0.973	0.990	0.995	0.989	0.982	0.976
R-Loss	0.998	0.998	0.994	0.994	0.996	0.994	0.994	0.995
SR-Precision	0.985	0.981	0.985	0.952	0.997	0.992	0.986	0.992

the intuition, as One Error considers only one label and Coverage is recall-based and takes the set of true labels into account. Recalling that MAP-Ex lies a high value on the highest ranked concept, the correlation of One Error and MAP-Ex is consequently (see Subplot i). However, it is astonishing that MAP-Ex still has a reasonable correlation to those measures that consider the complete set of labels.

9.4.2. Results of the stability experiment

The results of the stability experiment are presented in Table 9.3. For this experiment, the binary ground truth annotations were randomly flipped from zero to one and vice versa for 1%, 2%, 5%, and 10% of the test set. This refers to a maximum of 68,900 flipped labels in 2009 and 93,000 flipped labels in 2010. The overall ranking of runs is computed with each measure based on the original ground truth and regarding the four altered ground truths. Then, the rank correlations between the resulting ranking are calculated in two configurations. The first considers the correlation between the ground truth to all other stages of noise. The second configuration bases the rank correlation on the rankings of two noise stages.

Interestingly, the rank correlation is very high for all measures and presents stable rankings, even with 10% noise in the ground truth. The measure Coverage shows the highest influence of noise on the correlations. Compared to the altered ground truth with 10% noise, the correlation goes down to 0.91 and 0.81 in 2009 and 2010, respectively. Especially the difference between the 2009 and 2010 runs is apparent. Coverage computes how far one has to traverse the list of concepts in order to find all relevant labels. As the list of concepts in 2010 is 43% longer, the stability decreases with increasing noise. The measure SR-Precision shows the opposite behaviour. While it is generally less influenced by the noise with a correlation coefficient of 0.91 and 0.97 compared to the altered ground truth with 10% noise, it shows higher correlations for the 2010 runs. As

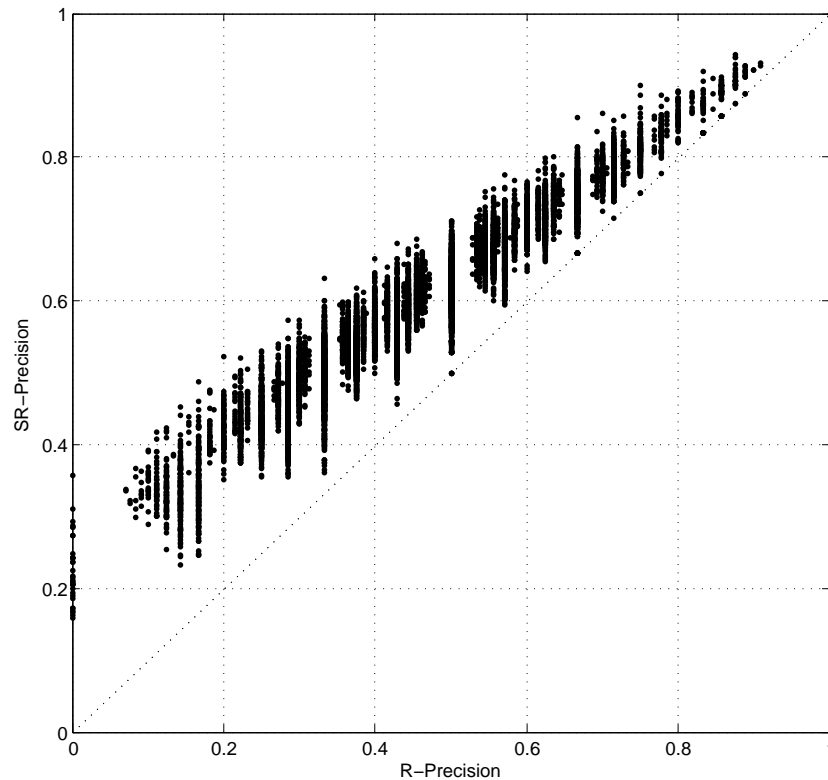


Figure 9.4.: Comparison of image scores assigned by R-Precision and SR-Precision for one run

the list of visual concepts is longer in 2010, there are more possibilities how concepts can be matched according to a high semantic relatedness. All other measures are only slightly influenced and show a very insensitive behaviour to the introduction of noise. While a general stability of evaluation measures is a desired property, these results do not discriminate well between the different measures. Further work has to show whether the stability of performance measures should be investigated in a different setting, and why they behave this insensitively to noise.

9.4.3. Comparison of performance scores between SR-Precision and R-Precision

Results of the correlation experiments have shown the high correlation between SR-Precision and R-Precision. The systems are almost ranked equally with an average correlation coefficient of 0.97 for the 2009 and 2010 runs. This poses the question of whether there are any benefits in incorporating the semantic relatedness among concepts into the performance assessment. The results on system ranking suggest that this is not necessary and it is better to use the more simple and already established R-Precision measure.

The benefits of incorporating semantics into the performance assessment come into play when analysing the variations in effectiveness score assignment of both measures to single images. Figure 9.4 depicts the effectiveness scores assigned to all 13,000 image predictions of one run which was submitted to the 2009 Photo Annotation task. The scores assigned by R-Precision are plotted on the x-axis, while the scores assigned by SR-Precision are plotted on the y-axis. It can be clearly seen that SR-Precision assigns a range of effectiveness scores with a maximum deviation of about +0.25 for badly annotated images and of about +0.1 for well annotated images in comparison to the R-Precision measure. All sample points plotted in a straight line in parallel to the y-axis denote images which are assessed as annotated equally well by the R-Precision measure. For all these images the predicted concept set has the same number of misclassified concepts in

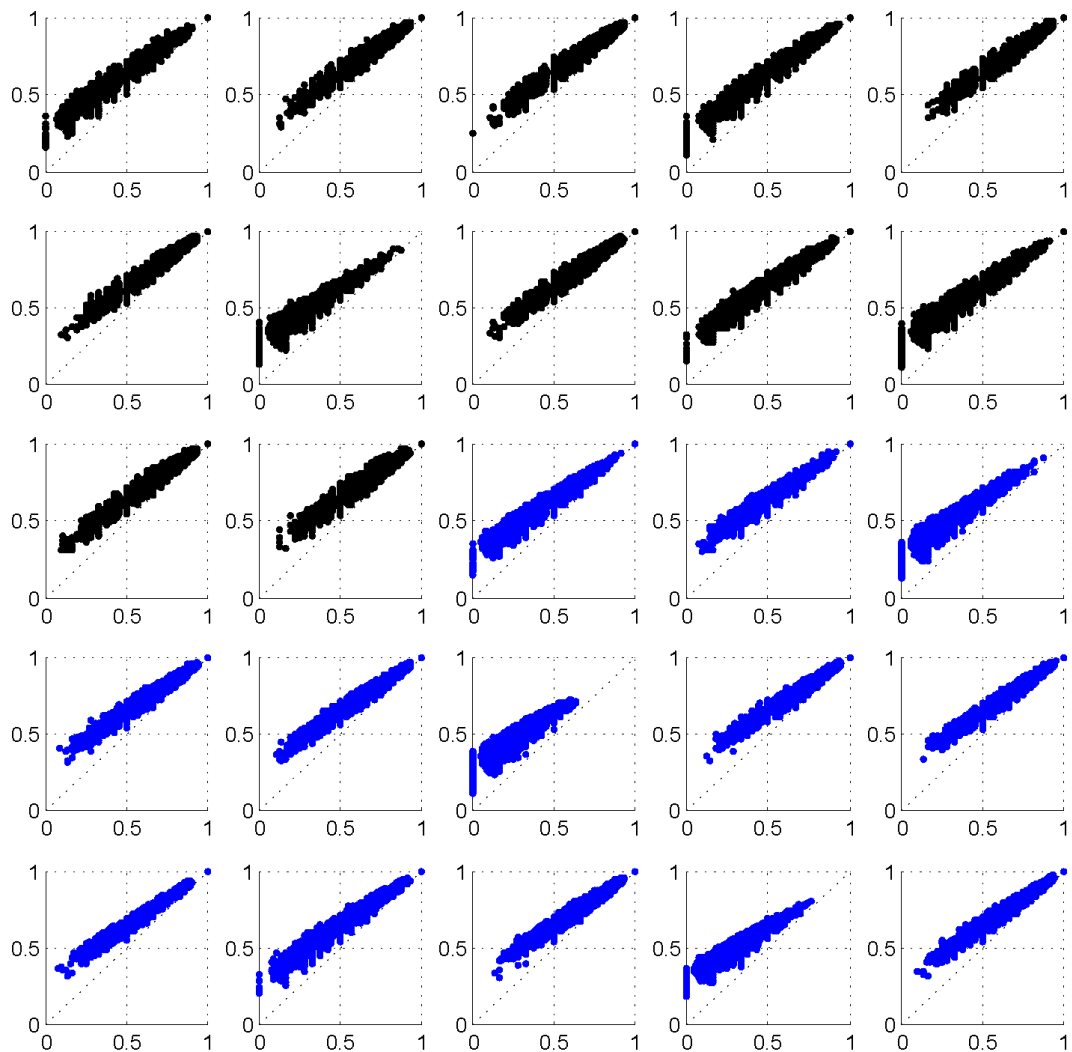


Figure 9.5.: Comparison of image scores for R-Precision and SR-Precision. The plots in black and blue correspond to runs from the Photo Annotation task in 2009 and 2010, respectively. The R-Precision is plotted on the x-axis in comparison to the SR-Precision on the y-axis

proportion to the LC of the particular ground truth set. However, the SR-Precision measure considers the semantic relation of predicted and ground truth concept sets. This enables the SR-Precision measure to distinguish among moderately well annotated images and badly annotated images. The lower extremum equals the score that is also assigned by R-Precision. Figure 9.5 demonstrates the distributions for other submissions to the 2009 and 2010 Photo Annotation tasks. The plot on top left depicts the same run as in Figure 9.4 on a smaller scale. The other plots in black refer to other runs from different teams of the 2009 task, while the ones in blue refer to runs from different teams submitted to the 2010 task. It is clearly visible that the illustrated pattern holds for all submissions.

9.4.4. Discussion

The experimental analysis showed the high correlation of SR-Precision to R-Precision which arises from a similar definition. While the R-Precision considers binary scores for misclassification, the

strengths of the SR-Precision come into play when dealing with misclassifications. SR-Precision is able to distinguish between predictions with an equal number of misclassifications by estimating the semantic relatedness of concepts in the prediction and ground truth. This provides the opportunity to differentiate between semantically close and unrelated concept predictions. The semantic relatedness measure FTS was chosen as a means to determine misclassification costs, as this measure has proven to cover the understanding of humans about semantic relatedness best. It therefore can be seen as an automated measure which judges according to human intuition.

The definition of SR-Precision further fits to general requirements on performance measures as defined in King (2003) (see Chapter 4.6). It reaches its highest value with perfect quality and its lowest value only for the worst possible quality. It is monotonic in the sense that a better quality goes with a higher score. Further, the SR-Precision is reliable in the sense that it exhibits little variance for equivalent inputs; it is automated and cheap to set up and apply.

Finally, King poses the requirement that a metric should be clear and intuitive and correlate well with human judgements. This requirement will be analysed in the following experiment. While I have proven that the FTS is adequate to determine relatedness of visual concepts in accordance to human opinion, the proposed user model has not been validated with respect to the user needs of image searchers. Especially, the question of whether the user experiences a benefit from the differentiation into moderately and worse annotated images based on the predicted concept set has to be investigated.

9.5. How close is the user model to real user needs?

In the following, the results of a survey on image search are described. The survey was conducted as a crowdsourcing experiment at MTurk. A total of 20 turkers each answered the same HIT. The questions of the HIT concentrate on the verification of the requirements stated in the user model. Especially, Requirement 2 about the benefits of a fine-grained assessment is considered.

The basic idea was to provide the turkers with the concepts sets predicted by two systems of the Photo Annotation task and ask them which concept set better fits to a given ground truth. However, manually comparing such a large number of concepts is confusing and annoying. Therefore, the submissions to the Photo Annotation task and the ground truth were reduced to 21 concepts. For two runs, 10 images were chosen that each got equal effectiveness scores in terms of R-Precision, but exhibit a difference in scores in terms of SR-Precision. These images are depicted in Table 9.4. For each image, the concept predictions of both submissions are visualised. The concepts written in blue correspond to correctly predicted concepts (TPs), the concepts written in red determine concepts that were predicted, but not contained in the ground truth (FPs), and the concepts printed in black are concepts that were missed by the runs (FNs). Images 1-5 contain one misclassification, images 6-9 include each two misclassifications, and image 10 was misclassified in three concepts. The turkers were asked to compare the red concepts with the black concepts and decide which system suggested better concepts.

Results are shown in Table 9.5. Except for image 10, the majority of turkers always preferred the concept set which also scored higher in terms of SR-Precision. This decision was independent of the absolute value of the SR-Precision score. For some of the images, such as the images 1,2,4,5, and 6, this decision was clearly performed. For others, such as the images 3 and 8, the turkers were more undecided. Image 10 accounts for the greatest number of misclassifications in this experiment. This image was only assessed by 9 out of 20 turkers. Several turkers could not distinguish between the concept sets and voted both sets as being equally well. It seems as if the human attention, at least in an uncontrolled environment which is the case in crowdsourcing experiments, cannot concentrate and compare this amount of misclassifications. Many turkers decided not to answer at all, while some chose the compromise. However, all in all, this experiment

Table 9.4.: Comparison of concept sets used in the MTurk survey









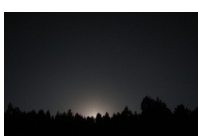
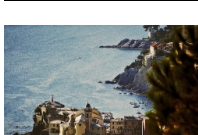
#	Image	System 1	System 2
1		Sky, Day, Single Person	Sky, Day, Citylife
2		Landscape-Nature, Sky, Clouds, Water, Sea, Day, Mountains, Plants	Landscape-Nature, Sky, Clouds, Water, Sea, Day, Mountains, Lake
3		Citylife, Day, Single Person, Animals, Small Group of People	Citylife, Day, Single Person, Animals, Small Group of People
4		Plants, Trees, Day, Flowers	Plants, Trees, Day, Animals
5		Landscape-Nature, Plants, Trees, Sky, Water, Day, Clouds	Landscape-Nature, Plants, Trees, Sky, Water, Day, Citylife
6		Plants, Trees, Day, Portrait, Small Group of People, Night, Single Person	Plants, Trees, Day, Portrait, Small Group of People, Sky, Single Person
7		Landscape-Nature, Plants, Trees, Sky, Day, Portrait, Single Person, Citylife, Water	Landscape-Nature, Plants, Trees, Sky, Day, Portrait, Single Person, Citylife, Small Group of People
8		Landscape-Nature, Plants, Sky, Day, Citylife, Single Person	Landscape-Nature, Plants, Sky, Day, Clouds, Single Person
9		Landscape-Nature, Plants, Trees, Sky, Night, Day, Sunset or Sunrise	Landscape-Nature, Plants, Trees, Sky, Night, Day, Single Person
10		Landscape-Nature, Plants, Trees, Water, Sea, Mountains, Day, red, Sky, Animals	Landscape-Nature, Plants, Trees, Water, Sea, Mountains, Day, red, Sky, Clouds

Table 9.5.: Effectiveness scores of the SR-Precision measure in comparison to the votes of the turkers on what concept set fits best to an image. The numbers printed in bold denote the better SR-Precision score or the greater number of votes.

Image	# FPs	SR-Precision system 1	SR-Precision system 2	# votes system 1	# votes system 2	# votes both
1	1	0.694	0.638	16	2	2
2	1	0.922	0.893	12	6	2
3	1	0.803	0.843	8	9	3
4	1	0.807	0.775	17	3	0
5	1	0.911	0.877	14	4	2
6	2	0.713	0.753	3	12	5
7	2	0.797	0.810	7	11	2
8	2	0.657	0.690	7	10	3
9	2	0.708	0.752	7	12	1
10	3	0.758	0.730	0	5	4

showed that human subjects are able to differentiate among different qualities of concept sets and that in most cases one annotation system is the clear winner in annotation performance. This winner is in 90% in line with the SR-Precision score.

Finally, the turkers should answer the question on what they consider most important when they are searching for images depicting several objects, and what their expectations about the results are. This question was intentionally posed vaguely in order not to lead the answers to a special direction. The comments from the turkers show that some did not know what to answer (e.g., "Im doing my best to help and to earn... Thanks!"). Some of the answers are presented in Table 9.6 in the wording that was used in the MTurk survey, except the correction of typos. The answers can be grouped into four types of expectations: find correct matches, find partly correct matches, no expectations, and other expectations. Most turkers require that all searched objects are visible in the image. Some even require that these objects are in the focus or fill the biggest part of the image. Two turkers are satisfied if at least a part of the searched objects is present. Nobody directly states that he is satisfied with semantically related objects or that he favours semantically related objects over unrelated ones. On the other hand, the answers illustrate the need of several, respectively, all concepts to be present in the images. This indirectly relates to Requirement 1 of the user model. If users judge the annotation performance by the presence of all concepts per image, the system-based evaluation should judge quality in the same way.

However, these answers have to be interpreted with care. The turkers were influenced by the questions they answered before and the experiment was conducted in an uncontrolled environment. To my point of view a thorough validation of the user model can just be performed by analysing the interaction of human subjects with concept-based image search systems, which is best conducted in a controlled study setup. Then, the perceived benefit, if any, can be assessed. Regrettably, such as user study was out of scope of this thesis.

9.6. Summary

In this chapter, I introduced the SR-Precision measure which considers rank-based predictions for example-based evaluation of annotation approaches. The user model for concept-based retrieval was extended and related to measures for ranked predictions. None of the existing measures

Table 9.6.: Expectations of turkers on image retrieval results

find correct matches	find partly correct matches	nothing	others
<ul style="list-style-type: none"> * the image should contain all things searched for * I expect the results to be relevant * when we search for an image, it should contain significant objects at least 1/2 of the total area. * the objects that are mentioned and the clarity. * most important are all items being used in image * they all match * images that include the searched objects * accuracy * Expect to see all the objects I asked for. Don't expect that the majority will be just what I want. 	<ul style="list-style-type: none"> * I expect the system at least shows two of the elements I'm searching for * The first few keywords and then having the most keywords that match. 	<ul style="list-style-type: none"> * no expectations! 	<ul style="list-style-type: none"> * the focus of the picture * Location and focus of objects are very important. * If I were searching for scenery I would prefer pictures with no people and only "nature" pictures.

are able to satisfy all requirements of the user model; this justifies the introduction of a novel performance measure. The SR-Precision fulfils all three requirements of the user model: it regards the complete ground truth set of labels, it assigns fine-grained scores in case of misclassifications, and it works on ranked predictions. This omits the need to binarize classifier outputs.

Three experiments were conducted in order to relate the SR-Precision measure to traditional example-based performance measures for ranked predictions. First, it was compared to five measures in extensive image and system ranking correlation experiments. Second, the stability of the different measures to artificially introduced noise in the ground truth was assessed. In the third experiment, the image assessment was directly compared to the image score predictions of the R-Precision measure. Results demonstrate the benefits of the SR-Precision to differentiate among varying qualities of image predictions, while retaining discriminative ranking power and stability characteristics in a high correlation to the R-Precision measure at the same time.

A MTurk experiment was conducted in order to verify if the benefit of differentiating image predictions in a fine-grained fashion is also perceived by human subjects. Results are promising, as in 90% the votes of the turkers are in line with the SR-Precision score. However, the interpretability of this experiment is limited and cannot replace an extensive user study in a controlled environment.

The user model on concept-based image search has shown the strong need for example-based performance measures in image annotation assessment. Depending on the system setup predictions might be better evaluated with a label set performance measure or with a performance measure working on ranked predictions. Both dimensions of the prediction format have been tackled in this thesis. In the need of a label set measure, the OS-FTS can be applied. For ranked predictions, the SR-Precision is a good choice for a fine-grained evaluation.

10. Image annotation evaluation in ImageCLEF



This chapter provides an overview of the photo annotation evaluation activities in ImageCLEF. The goals, achievements, and evolvments of object and concept recognition for image retrieval are presented. The first automatic annotation challenge was posed in 2006, but a special focus in this chapter lies on the ImageCLEF tasks in 2009 and 2010, and on the International Conference on Pattern Recognition (ICPR) task in 2010, as these tasks were proposed, organised, and analysed by the author. Especially, this chapter unites the achievements on test collections and evaluation measures detailed in this thesis. The OS and Ontology Score with Flickr Context Similarity (OS-FCS) measures, as introduced in Chapter 7 and 8, were applied and tested in a real-world application scenario. After a short introduction in Section 10.1, the chapter continues with the VCDT in general and the evaluation objectives of the three cycles in Section 10.2. Next, Section 10.3 introduces the test collections, including the relevance assessment process. Further, the evaluation methodology is detailed in Section 10.4, followed by the submission format in Section 10.5 and the participating groups in Section 10.6. The results are illustrated in Section 10.7. Section 10.8 presents the evolvment over the years, followed by an outlook on ImageCLEF 2011 in Section 10.9. Finally, Section 10.10 closes the chapter with a short summary.

10.1. Motivation

The steadily increasing amount of multimedia data poses challenging questions on how to index, visualise, organise, navigate, or structure multimedia information. Many different approaches have been proposed in the research community, but their benefit is often not clear, as they were evaluated on different datasets with different evaluation measures. Evaluation campaigns aim at establishing an objective comparison between the performance of different approaches by posing

well-defined tasks including datasets, topics, and measures. Evaluation initiatives in multimedia became popular with the text-based evaluations of TREC, the video analysis evaluation of TRECVID, and the multi-modal, cross-lingual evaluation efforts of CLEF. Since 2003, ImageCLEF has been a part of the CLEF evaluation initiative. It focuses on the evaluation of multi-modal image retrieval approaches in the consumer and medical domain. For a general introduction into benchmarking in MIR, the reader is referred to Chapter 3.4.

10.2. Visual concept detection and annotation task

In 2006, ImageCLEF added an “Automatic annotation task for general photographs” which, over the years, evolved from a flat image classification task into an object retrieval task (2007), then into a hierarchical concept annotation task (2008-2009), and, finally to a multi-modal hierarchical annotation task utilising user-generated tags (2010). It has developed into an inherent part of the annual ImageCLEF evaluation cycle with interactions to other tasks. As the task names indicate, the focus of the task changed over the years, but its objective has always been to analyse the content of images based on their visual appearance.

Evaluation campaigns for object detection (Everingham et al. (2006, 2010)), content-based image retrieval (Clough et al. (2005)), and image classification (Moellic and Fluhr (2006)) have established themselves since 2005. Smaller evaluation events such as the ImageNet Large Scale Image Recognition task¹, started recently. Although these evaluation initiatives have a certain overlap with the tasks described in this chapter, ImageCLEF has always focused on multi-modal analysis and the integration of detection technologies into actual retrieval systems. For example, in 2006 and 2007, the generalisation of object recognition algorithms across different databases was tested. This scenario denies the frequent assumption that for training and test, the same database with similar annotation characteristics is present. In 2008, the participants were provided with a taxonomy, and in 2009, with an ontology as additional knowledge sources. These knowledge sources structured the visual concepts into sub- and super-classes and specified relations and restrictions. This textual information was available to enhance the visual analysis algorithms and to validate the output of the classifiers. In 2010, the task was split into three configurations, namely visual, textual, and multi-modal, as in that year, user-generated tags were additionally provided and the task could be solved purely textually for the first time. Further, ImageCLEF proposed tasks in form of a contest on an additional venue in 2010. The following sections will mainly focus on the task from 2009 and the two from 2010, as the author was responsible for the organisation from 2009 onwards. Details on the tasks from 2006-2008 can be found in Nowak et al. (2010).

10.2.1. Evaluation objectives in 2009

The VCDT 2009 poses the challenge of multi-label classification in consumer photos with the help of ontology knowledge. The participants were provided with a training set of annotated Flickr photos and asked to automatically annotate a test set with a number of visual concepts. The visual concepts are organised in the PTO as introduced in Chapter 5.5. The PTO was additionally provided so that the participants could make use of the hierarchical order and the relations between concepts for solving the annotation task. The VCDT mainly addresses two issues:

1. Can image classifiers scale to the large amount of concepts and data?
2. Can an ontology (hierarchy and relations) help in large scale annotations?

¹<http://www.image-net.org/challenges/LSVRC/2010/>, last accessed 20.02.2011

The challenge of the annotation task consists in coping with the unbalanced amount of data per concept, with the subjectivity of the presence of some concepts, as well as with the diversity of images belonging to the same concept class. Approaches that try to adopt the ontology in the learning process are appreciated, as the question of whether incorporating real-world knowledge leads to superior results in contrast to applying purely machine learning approaches is still open.

10.2.2. Evaluation objectives at the ICPR contest 2010

In 2010, ImageCLEF performed an additional benchmarking event at the ICPR. This event was out of the usual evaluation cycle with a tough timetable. At the ICPR contest, the VCDT posed a similar problem as in ImageCLEF 2009. Again, participants were asked to automatically annotate a set of Flickr photos with visual concepts. This time, an additional validation set was provided and the number of photos in test and training set changed in contrast to 2009. However, the research challenge remained the same.

10.2.3. Evaluation objectives in 2010

In ImageCLEF 2010, the VCDT included user-generated metadata as additional textual resource and a differentiation was made between three configurations:

1. Automatic annotation with content-based visual information of the images.
2. Automatic annotation with Flickr user tags and EXIF metadata in a purely textual scenario.
3. Multi-modal approaches that consider both visual and textual information, such as Flickr user tags or EXIF information.

In all cases, the participants of the task were asked to annotate the photos of the test set with a predefined set of concepts, allowing for an automated evaluation and comparison of the different approaches. The number of visual concepts was substantially extended. The focus of the task lies on the comparison of the strengths and limitations of the different approaches:

- Do multi-modal approaches outperform text-only or visual-only approaches?
- Which approaches are best for which kind of concepts?
- Can image classifiers scale to the large number of concepts and data?

Further, the task challenges the participants to deal with an unbalanced number of annotations per image, an unbalanced number of images per concept, the subjectivity of concepts such as *boring*, *cute*, or *fancy*, and the diversity of images belonging to the same concept. Additionally, the textual runs have to cope with a small number of images without EXIF data and/or Flickr user tags.

10.3. Test collection

All three cycles of the VCDT in 2009 and 2010 made use of a subset of the MIR Flickr 25,000 image dataset of Huiskes and Lew (2008). This collection contains 25,000 photos from the Flickr platform that were collected based on the interestingness rating of the community and the creative commons copyright of the images. The set used for the ImageCLEF cycles contains 18,000 Flickr photos in total (see Table 10.1). In all evaluation cycles, the VCDT test collection was fully assessed with relevance judgements. The ImageCLEF 2009 task used 5,000 images as

Table 10.1.: Number of photos and concepts used in the ImageCLEF cycles.

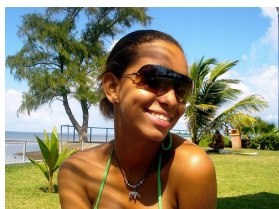
	# Training	# Validation	# Test	# Visual Concepts
ImageCLEF 2009	5,000	-	13,000	53
ICPR 2010	5,000	3,000	10,000	53
ImageCLEF 2010	8,000	-	10,000	93

training set and 13,000 images as test set, while in the ICPR contest, 5,000 images for training, 3,000 images for validation, and 10,000 for the test were utilised. These photos were all manually annotated with 53 visual concepts (see Table A.1). The concepts resulted from the studies described in Chapter 5 and are organised in the PTO.

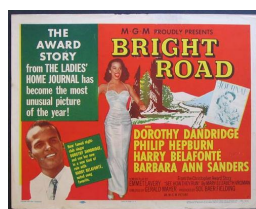
In ImageCLEF 2010, 8,000 images were used for training, and 10,000 for the test. Basically, the training and validation set of the ICPR benchmark were used as training set for 2010 and the test set remained the same. The number of visual concepts was substantially extended to 93 concepts (see Table A.2). 52 of the 53 former concepts were reused. In contrast to the previous annotations, the new annotations were obtained with a crowdsourcing approach that utilises MTurk. The study on annotation quality obtained by assessors from MTurk in Chapter 6 showed that it is reasonable to outsource the relevance assessment. As textual resource, the MIR Flickr collection supplies all original tag data provided by the Flickr users (noted as Flickr user tags). In the collection, there are 1386 tags which occur in at least 20 images, with an average total number of 8.94 tags per image. These Flickr user tags are made available for the textual and multi-modal approaches. For most of the photos, the EXIF data is included and may be used.

10.3.1. Relevance assessment in 2009

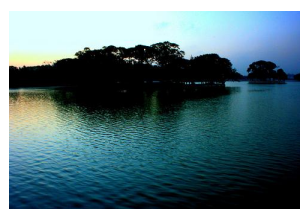
The annotation of 18,000 Flickr photos with 53 concepts for the 2009 and 2010 evaluation cycles was conducted by 43 persons of the Fraunhofer IDMT research institute. The number of photos that were annotated by one person varied between 30 and 2,500 images. All annotators were provided with a definition of the concepts and example images, with the goal of allowing a consistent annotation among the large number of persons. Some of the concepts exclude each other, while others can be depicted simultaneously. After this first step, the annotations were validated. Due to the number of assessors, the number of photos, and the ambiguity of some image content, the annotations were not consistent throughout the database. Three persons performed a validation by screening only those photos that (a) were annotated with concept



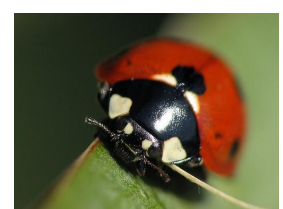
Family-Friends, Sky, Summer, Outdoor, Trees, Clouds, Day, Sunny, Portrait, Single Person, Neutral Illumination, No Blur



Canvas, No Blur, Neutral Illumination, Small Group, No Visual Season, No Visual Place, No Visual Time



Landscape, Outdoor, Water, Trees, Sky, Day, Overexposed, No Blur, No Persons, No Visual Season



Plants, Outdoor, Partly Blurred, Macro, Animals, No Visual Time, Neutral Illumination, No Persons, Summer, Aesthetic Impression

Figure 10.1.: Example images from the visual concept detection task in 2009.

X and (b) that were not annotated with concept X. In the first case, they had to delete all annotations for concepts that were not depicted in the photo (FPs). In the second case, the goal was to find the photos with a missing annotation for concept X, but an obvious appearance of this concept (FNs). A subset of 100 photos was assessed by 11 different persons. These annotations are used to determine the annotator agreement included in the OS computation. For each photo and each concept, the annotation of the majority of annotators was regarded as correct, and the percentage of annotators that annotated correctly is utilised as agreement factor. Figure 10.1 shows four example photos and the corresponding annotations. The number of annotations per photo varied substantially; for example, the annotations of the photos in Figure 10.1 range from 7 to 12 labels per photo.

10.3.2. Relevance assessment in 2010

For ImageCLEF 2010, 41 new concepts were annotated with a crowdsourcing approach using MTurk (see Chapter 4.2 for a description of the service of Amazon). Before the annotations of the ImageCLEF 2010 tasks were acquired, a pre-study was performed in order to investigate whether annotations from non-experts are reliable enough to be used in an evaluation benchmark. The results were very promising and encouraged me to adapt this service for the 2010 task. Details of the pre-study can be found in Chapter 6.

In total, four different HIT templates at MTurk were designed. These made use of pre-knowledge obtained from the 2009 annotations. In the first three groups, each HIT was rewarded with 0.01\$, while 0.03\$ were paid per HIT in the last group. In total, annotations were obtained three times for all concepts and images. Afterwards, the final annotations are built from the majority vote of these three opinions.

1. Vehicles

A number of images was already annotated with the concept *vehicle*. These images were further classified concerning the concepts *car*, *bicycle*, *ship*, *train*, *airplane*, and *skateboard*. The corresponding survey with guidelines is illustrated in Figure 10.2.

2. Animals

Several images were annotated with the concept *animals* in 2009. The turkers further classified these images in the categories *dog*, *cat*, *bird*, *horse*, *fish*, and *insect*.

3. Persons

The dataset contains images that were annotated with a person concept (*single person*, *small group*, or *big group* of persons). These images were further classified with human attributes such as *female*, *male*, *baby*, *child*, *teenager*, *adult*, and *old person*.

4. General annotations

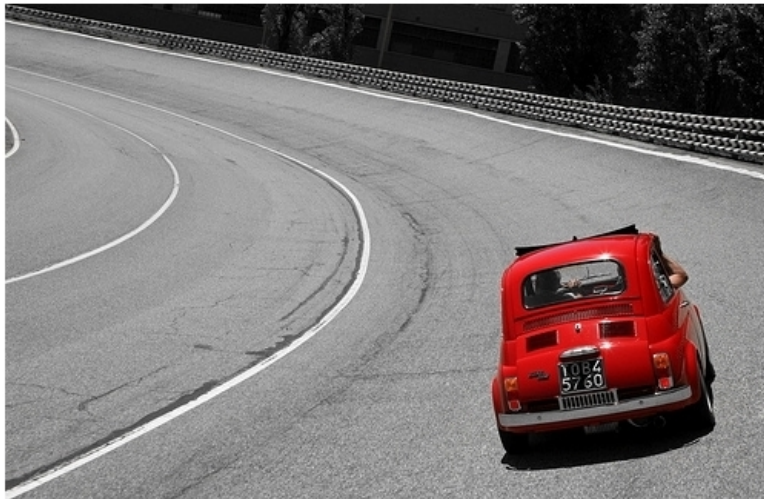
For the last 22 new concepts, no prior information could be used. Therefore, the concepts were annotated in all 18,000 photos. The HIT was designed as a survey with 6 questions, aiming to annotate the categories *content elements* (architecture, street, church, bridge, park/garden, rain, toy, musical instrument, and shadow), *persons* (bodypart), *events* (travel, work, birthday), *image representation* (visual arts, graffiti, painting), *impression* (artificial, natural, technical, abstract), and *feelings* (boring, cute).

Annotate this image

Guidelines:

- Please choose all applicable concepts for an image, but at least one.
- In case no definition is matching, use the text area to classify the vehicle.
- In case there are several vehicles shown, please answer for all of them.
- Please note that vehicles in paintings are also considered to be vehicles and should be classified.

Image:



Vehicles

1. Which vehicle(s) can you see at the photo (select all matching answers)

car bicycle ship / boat train airplane skateboard

Another vehicle? There is no vehicle shown.

Figure 10.2.: MTurk HIT template for the annotation of specific vehicle concepts.

10.4. Evaluation measures

The evaluation follows the concept-based and example-based evaluation paradigms, as introduced in Chapter 3.3. The submissions to the VCDT task in 2009 and at ICPR 2010 were evaluated with the same concept-based and example-based performance measures. The concept-based evaluation was performed with the EER and AUC, while the example-based evaluation considered the OS (see Chapter 7.3). In 2010, the concept-based evaluation used the iAP. This measure showed better characteristics than the EER and AUC (see Chapter 7.5). The EER and AUC scores were additionally calculated to be able to compare the results to those of the previous year. For the example-based evaluation, the example-based F-Measure (F_{eb}) was applied. Second, the extended OS-FCS served as additional evaluation measure (see Chapter 8).

10.5. Submission

The participants submitted the results for all photos in one text file that contains the photo ID as first entry per row, followed by a floating point value between 0 and 1 for each concept. The floating point values are regarded as confidences while computing the confidence-based evaluation

Table 10.2.: Participation in the concept detection task over the years. Please note that the maximum number of runs was restricted to five per group.

	# Groups registered	# Groups participated	# runs
2009	34	19	73
ICPR 2010	17	12	44
2010	41	17	63

measures such as EER, AUC, and AP. For the binary measures such as OS, OS-FCS, and F_{eb} a decision about the presence or absence of the concepts is needed. In 2009, the participants were asked to provide a threshold with which a binary annotation could be obtained from the confidence values. At ICPR 2010, this was further relaxed so that the participants could provide an individual threshold per concept. Finally, in ImageCLEF 2010, the participants had the possibility to threshold each concept for each image individually. Therefore, after the confidence values for all photos, the submission text file contained binary annotations for each photo and concept. All groups had to submit a short description of their runs. In ImageCLEF 2010, this included which configuration they chose (annotation with visual information only (abbreviated with “V”), annotation with textual information only (“T”), or annotation with multi-modal information (“M”).

10.6. Participation

In the three evaluation cycles, 29 international research groups participated in total. A subset of 14 research groups took the challenge several times. The participation of the groups is summarised in Table 10.2. All participants needed to sign a license agreement in order to access the data and annotations. As is visible from Table 10.2, research groups often require access to the test collection, but do not submit results. All participating groups are listed together with the group acronyms and the references to their approaches in the following:

- **apexlab:** (Nowak and Dunker (2010)): Shanghai Jiaotong University, Shanghai, China
- **AVEIR:** (Glotin et al. (2009); Nowak (2010b)): joint consortium of the four groups: Telecom ParisTech, LSIS, MRIM-LIG and UPMC
- **budapest / sztaki:** (Daróczy et al. (2010), Daróczy et al. (2010)): Data Mining and Web search Research Group, Informatics Laboratory, Computer and Automation Research Institute, Hungarian Academy of Sciences, Hungary
- **CEA LIST:** (Nowak and Dunker (2010); Nowak and Huiskes (2010)): Lab of Applied Research on Software-Intensive Technologies of the CEA, France
- **CVSSPret:** (Tahir et al. (2010)): Centre for Vision, Speech and Signal Processing, Department of Electronics, University of Surrey, UK
- **DCU:** (Li et al. (2010)): Dublin City University, Dublin, Ireland
- **FIRST:** (Binder and Kawanabe (2010)): Fraunhofer FIRST, Berlin, Germany
- **HHI:** (Mbanya et al. (2010)): Fraunhofer HHI, Berlin, Germany
- **I2R:** (Ngiam and Goh (2010)): IPAL French-Singaporean Joint Lab of the Institute for Infocomm Research, Singapore

- **IAM:** (Hare and Lewis (2009)): Intelligence Agents Multimedia Group of the University Southampton, UK
- **IJS:** (Nowak (2010b); Dimitrovski et al. (2010)) Team of Jožef Stefan Institute, Slovenia and Department of Computer Science, Macedonia
- **INAOE TIA:** (Escalante et al. (2009)): TIA Research Group, Computer Science Department, National Institute of Astrophysics, Optics and Electronics, Tonantzintla, Mexico
- **INSUNHIT:** (Zhang et al. (2010)): Harbin Institute of Technology, Harbin, China
- **ISIS:** (van de Sande et al. (2010); Nowak (2010b); van de Sande and Gevers (2010)): Intelligent Systems Lab of the University of Amsterdam, The Netherlands
- **ITI:** (Nowak (2010b)) The team of the Institute of Computer Technology, Polytechnic University of Valencia, Spain
- **Kameyama:** (Sarin and Kameyama (2009)): Graduate School of Global Information and Telecommunication Studies, Waseda University, Japan
- **LEAR:** (Douze et al. (2009); Mensink et al. (2010)): LEAR team of INRIA, France
- **LSIS:** (Dumont et al. (2010); Nowak (2010b); Paris and Glotin (2010)): Laboratory of Information Science and Systems, France
- **MEIJI:** (Motohashi et al. (2010)): Meiji University, Kanagawa, Japan
- **MLKD:** (Nowak and Huiskes (2010)): Aristotle University, Thessaloniki, Greece
- **MMIS:** (Llorente et al. (2010); Nowak (2010b)): Knowledge Media Institute, Open University, Milton Keynes, UK
- **MRIM-LIG:** (Pham et al. (2010); Nowak (2010b); Batal and Mulhem (2010)): Multimedia Information Modelling and Retrieval group at the Laboratoire Informatique de Grenoble, Grenoble University, France
- **Romania:** (Rasche and Vertan (2010)): University of Bucharest, Bucharest, Romania
- **Telecom ParisTech|CNRS:** (Ferecatu and Sahbi (2009); Nowak (2010b); Sahbi and Li (2010)): Institut TELECOM, TELECOM ParisTech, Paris, France
- **TRS2008:** (Nowak (2010b)) Beijing Information Science and Technology University, China
- **UAIC:** (Iftene et al. (2010); Nowak (2010b)): Faculty of Computer Science of Alexandru Ioan Cuza University, Romania
- **UPMC/LIP6:** (Fakeri-Tabrizi et al. (2010); Nowak (2010b); Fakeri-Tabrizi et al. (2010)): University Pierre et Marie Curie, Paris, France
- **Wroclaw Uni:** (Nowak and Dunker (2010), Stanek and Maier (2010)): Wroclaw University of Technology, Wroclaw, Poland
- **XRCE:** (Ah-Pine et al. (2009); Mensink et al. (2010)): Textual and Visual Pattern Analysis group from the Xerox Research Center Europe, France

Table 10.3.: Summary of the results for the VCDT in 2009. The table shows the EER and AUC performance for the best run per group ranked by EER for the concept-based evaluation, and the performance with the OS measure for the best run per group for the photo-based evaluation. Note that the best run for the EER measure is not necessarily the same run as the best run for the OS measure.

Group ID	Descriptor	Classifier	Rank	EER	AUC	Rank	OS
ISIS	colour SIFT	SVM	1	0.23	0.84	14	0.77
LEAR	BoW (global and local)	SVM / NN	5	0.25	0.82	12	0.77
I2R	global and local	SVM	7	0.25	0.81	2	0.81
FIRST	SIFT, colour	multiple kernel SVM	8	0.25	0.82	4	0.80
XRCE	BoW	sparse logistic regression	14	0.27	0.80	1	0.81
budapest	various global and local features	logistic regression	17	0.29	0.77	35	0.68
MMIS	colour, Tamura, Gabor	non-parametric density estimation	21	0.31	0.74	42	0.58
IAM	SIFT	cosine distance of visual terms	23	0.33	0.72	61	0.41
LSIS	various features	SVM(LDA*) / Visual Dictionary	24	0.33	0.72	49	0.51
UPMC	HSV histogram	SVM	33	0.37	0.67	58	0.44
MRIM	RGB histogram, SIFT, Gabor	SVM	34	0.38	0.64	28	0.72
AVEIR	various global and local features, text	SVM / Visual Dictionary / canonical correlation	41	0.44	0.55	50	0.50
Wroclaw Uni	various features	Multivariate Gaussian Model + NN	43	0.45	0.22	11	0.78
Kameyama	global and local	k -NN	47	0.45	0.16	7	0.80
UAIC	face detection, EXIF	NN + default values	54	0.48	0.11	32	0.69
apexlab	various features	k -NN	56	0.48	0.07	17	0.76
INAOE TIA	various global features	k -NN	57	0.49	0.10	20	0.74
Random	-	-	-	0.50	0.50	-	0.38
CEA LIST	global and local	Multiclass boosting	68	0.50	0.47	29	0.71
TELECOM	global, text features	Canonical Correlation Analysis + thresholds	72	0.53	0.46	65	0.39

10.7. Results

This section summarises the results of the three evaluation cycles, including a brief classification of the applied approaches. Detailed information on the approaches of the participants and the results can be found in the corresponding overview papers: Nowak and Dunker (2010), Nowak (2010b), Nowak and Huiskes (2010), and on the benchmark website (www.imageclef.org).

10.7.1. Results of the VCDT in ImageCLEF 2009

The best results per team of the VCDT in 2009 are listed in Table 10.3. In total, 73 runs were submitted by 19 groups. The team with the best concept-based results, ISIS, achieves an EER of 23% and an AUC of 84% on average for their best run. The next three teams in the ranking closely follow these results with an EER of about 25% and an AUC of 82% and 81%. The performance of the teams at the end of the list increases to 53% EER and decreases to 7% AUC. The evaluation per photo reveals scores in the range of 39% to 81% for the best run per group. The best results in terms of OS were achieved by the XRCE group with 81% annotation score over all photos.

Table 10.4.: Overview of concepts and results per concept in terms of the best EER and best AUC per concept and the name of the group which achieved these results. The results for the Photo Annotation task in 2009 are illustrated in the middle and those for ICPR 2010 on the right.

No.	Concept	Group	Best AUC 09	Best EER 09	Group	Best AUC ICPR 10	Best EER ICPR 10
0	Partylife	ISIS	0.84	0.24	CVSSPRet	0.87	0.22
1	Family-Friends	ISIS	0.83	0.24	CVSSPRet	0.86	0.22
2	Beach Holidays	ISIS	0.91	0.16	CVSSPRet	0.93	0.13
3	Building-Sights	ISIS	0.88	0.20	CVSSPRet	0.90	0.18
4	Snow	LEAR	0.85	0.22	CVSSPRet	0.89	0.19
5	Citylife	ISIS	0.83	0.24	CVSSPRet	0.85	0.22
6	Landscape	ISIS	0.95	0.13	CVSSPRet / ISIS	0.95	0.12
7	Sports	FIRST	0.72	0.33	CVSSPRet	0.78	0.29
8	Desert	ISIS	0.89	0.18	ISIS / CVSSPRet	0.92	0.18
9	Spring	FIRST	0.83	0.24	IJS	0.86	0.20
10	Summer	ISIS	0.81	0.25	CVSSPRet	0.83	0.23
11	Autumn	ISIS	0.86	0.21	ISIS	0.88	0.18
12	Winter	ISIS	0.84	0.23	CVSSPRet	0.88	0.21
13	No-Visual-Season	ISIS	0.80	0.26	CVSSPRet	0.82	0.25
14	Indoor	ISIS	0.83	0.25	CVSSPRet / ISIS	0.84	0.24
15	Outdoor	ISIS	0.90	0.19	ISIS	0.91	0.18
16	No-Visual-Place	ISIS	0.79	0.29	CVSSPRet	0.81	0.27
17	Plants	ISIS	0.88	0.21	CVSSPRet	0.90	0.18
18	Flowers	ISIS / FIRST	0.87	0.21	CVSSPRet	0.89	0.19
19	Trees	ISIS	0.90	0.18	CVSSPRet	0.92	0.16
20	Sky	ISIS	0.95	0.12	ISIS	0.96	0.10
21	Clouds	ISIS	0.96	0.11	ISIS	0.96	0.10
22	Water	ISIS	0.90	0.18	CVSSPRet	0.91	0.16
23	Lake	ISIS	0.90	0.17	CVSSPRet / ISIS	0.91	0.16
24	River	ISIS	0.90	0.17	CVSSPRet	0.93	0.14
25	Sea	ISIS	0.94	0.12	CVSSPRet	0.95	0.12
26	Mountains	ISIS	0.94	0.14	ISIS	0.95	0.12
27	Day	ISIS	0.85	0.24	CVSSPRet	0.87	0.22
28	Night	LEAR	0.91	0.17	IJS	0.92	0.16
29	No-Visual-Time	ISIS	0.84	0.25	CVSSPRet / ISIS	0.86	0.23
30	Sunny	LEAR / FIRST	0.77	0.30	CVSSPRet	0.81	0.27
31	Sunset-Sunrise	ISIS	0.96	0.11	ISIS / CVSSPRet	0.96	0.08
32	Canvas	I2R / XRCE	0.83	0.24	CVSSPRet	0.85	0.22
33	Still-Life	ISIS	0.83	0.25	CVSSPRet	0.86	0.22
34	Macro	ISIS	0.81	0.27	ISIS	0.84	0.24
35	Portrait	XRCE / ISIS	0.87	0.21	CVSSPRet	0.91	0.18
36	Overexposed	ISIS / UPMC	0.81	0.25	ISIS / CNRS	0.83	0.24
37	Underexposed	I2R	0.89	0.18	AVEIR / ITI	0.89	0.19
38	Neutral-Illumination	LEAR	0.80	0.26	IJS	0.81	0.26
39	Motion-Blur	ISIS	0.75	0.31	CVSSPRet / IJS	0.79	0.28
40	Out-of-focus	LEAR	0.82	0.25	CNRS / CVSSPRet	0.84	0.24
41	Partly-Blurred	LEAR	0.86	0.22	ISIS / CVSSPRet	0.87	0.21
42	No-Blur	LEAR	0.85	0.23	ISIS	0.86	0.22
43	Single-Person	ISIS / LEAR	0.80	0.27	CVSSPRet	0.83	0.25
44	Small-Group	ISIS	0.80	0.28	CVSSPRet	0.83	0.25
45	Big-Group	ISIS	0.88	0.20	CVSSPRet	0.91	0.17
46	No-Persons	ISIS	0.86	0.22	CVSSPRet / ISIS	0.87	0.21
47	Animals	ISIS	0.84	0.24	CVSSPRet	0.87	0.21
48	Food	ISIS	0.90	0.19	CVSSPRet	0.92	0.17
49	Vehicle	ISIS	0.83	0.24	CVSSPRet	0.86	0.23
50	Aesthetic-Impression	ISIS	0.66	0.39	CVSSPRet / ISIS	0.67	0.38
51	Overall-Quality	ISIS	0.66	0.39	CVSSPRet / ISIS	0.66	0.39
52	Fancy	LEAR / ISIS	0.59	0.44	ISIS / IJS	0.61	0.42

Table 10.5.: Results of the ICPR Photo Annotation task. The table lists the best results for each measure per group, the number of runs submitted and the descriptors and classifiers applied. It is sorted ascending due to the EER measure. Please note that the best run of a group for one measure is not necessarily the best run evaluated with the other measures.

Group ID	#	Descriptor	Classifier	EER	Rank	AUC	Rank	OS	Rank
CVSSPRet	5	various SIFT	spectral regression	0.21	1	0.86	1	0.69	5
ISIS	2	various SIFT	SVM	0.22	4	0.86	4	0.78	1
IJS	4	global and local	random forests	0.24	8	0.83	8	0.71	3
CNRS	5	global and local	SVM, boosting	0.28	12	0.79	12	0.42	23
AVEIR	4	global and local	SVM	0.29	17	0.79	17	0.56	12
MMIS	5	various features	non-parametric density estimation / MRF	0.31	19	0.76	19	0.50	17
LSIS	4	various features	SVM + reranking	0.31	21	0.75	21	0.51	16
UPMC	5	SIFT	SVM	0.34	28	0.72	29	0.40	28
ITI	5	local and global	NN	0.37	30	0.59	37	0.40	30
MRIM	3	colour, texture, feature points	SVM + Fusion	0.38	31	0.64	30	0.58	9
TRS2008	1	SIFT	SVM	0.42	34	0.62	33	0.33	38
UAIC	1	face detection, EXIF	NN + default values	0.48	38	0.14	43	0.68	6

In Table 10.4, the best results for each concept are illustrated in terms of EER and AUC over all runs submitted. All concepts could be detected at least with 44% EER and 58% AUC, but in average with an EER of 23% and an AUC of 84%. The majority of the concepts was classified best by the ISIS group. It is obvious that the aesthetic concepts (Aesthetic Impression, Overall Quality, and Fancy) are classified worst (EER greater than 38% and AUC less than 66%). This is not surprising due to the subjective nature of these concepts, which also made the relevance assessment process difficult. The best classified concepts are Clouds (AUC: 96%), Sunset-Sunrise (AUC: 95%), Sky (AUC: 95%), and Landscape-Nature (AUC: 94%).

Summarising, the groups that used local features such as SIFT achieved better results than the groups relying solely on global features. Most groups that investigated the concept hierarchy and analysed, for example, the correlations between the concepts, could achieve better results in the OS compared to the EER. The discriminative methods outperformed the generative and model-free ones.

10.7.2. Results of the VCDT at ICPR 2010

In the following, the results for the teams participating in the ICPR benchmark are discussed. Solutions from 12 research groups were submitted to the Photo Annotation task in a total of 44 run configurations. Table 10.5 summarises the performance for each group for the measures EER, AUC, and OS. The best results per team are listed together with a short characterisation of the adopted approach and the rank information of the run. The table is sorted ascending to the EER scores. Please note that the best run of a group for one measure is not necessarily the best run evaluated with the other measures.

For the concept-based evaluation, the best run achieved an EER of 21.4% and an AUC of 86% per concept on average (see Table 10.4). Two other groups got close results with an AUC score of 85.7% and 83.2%. The best annotation quality for a concept achieved by any run is on average 20.8% EER and 86.5% AUC. The concepts Sunset-Sunrise (AUC: 96.2%), Clouds (AUC: 96.2%), and Sky (AUC: 95.9%) are the easiest detectable concepts. Again, the worst concept

Table 10.6.: Summary of the results for the VCDT in 2010. The table shows the MiAP, F_{eb} and OS-FCS scores for the best run per group and indicates the configuration of the run.

Group ID	#	Rank	MiAP	Conf.	Rank	F_{eb}	Conf.	Rank	OS-FCS	Conf.
XRCE	5	1	0.46	M	5	0.66	M	1	0.66	M
LEAR	5	3	0.44	M	15	0.60	M	32	0.41	M
ISIS	5	5	0.41	V	1	0.68	V	10	0.60	V
HHI	5	16	0.35	V	8	0.63	V	3	0.64	V
IJS	4	20	0.33	V	18	0.60	V	12	0.60	V
MEIJI	5	23	0.33	M	23	0.57	M	30	0.43	M
CNRS	5	28	0.30	M	43	0.35	M	31	0.42	M
BPACAD	1	33	0.28	V	38	0.43	V	29	0.44	V
Romania	5	34	0.26	V	29	0.53	V	17	0.56	V
INSUNHIT	5	36	0.24	V	53	0.21	V	43	0.37	V
MLKD	3	37	0.24	M	49	0.26	T	42	0.38	M
LSIS	2	38	0.23	V	30	0.53	M	21	0.54	V
DCU	1	44	0.23	T	60	0.18	T	60	0.30	T
LIG	1	46	0.23	V	35	0.48	V	22	0.53	V
WROCLAW	5	50	0.19	V	34	0.48	V	41	0.38	V
UPMC	5	54	0.18	M	55	0.19	M	55	0.35	M
CEA-LIST	1	61	0.15	V	37	0.45	V	28	0.46	V

detection quality can be found for the concepts *Fancy* (AUC: 61.4%), *Overall Quality* (AUC: 66.1%), and *Aesthetic* (AUC: 67.1%). The example-based evaluation reveals that the best run is able to correctly annotate a photo with an average of 78.4%. Taking into account the best annotation quality per photo out of all runs, the photos can be annotated on average with 85.1% quality, ranging between 59.3% and 100%. The ranking of the runs is different than that of the concept-based measures.

To summarise, the best teams applied discriminative approaches with local features. Some used a fusion of local and global features. One model-free approach (UAIC) considering a combination of several methods could achieve good results in the OS.

10.7.3. Results of the VCDT in ImageCLEF 2010

In the Photo Annotation Task 2010, 17 groups participated with a total of 63 runs. In Table 10.6, the results for the evaluation independent of the applied configuration are illustrated for the best run of each group. The task could be solved best with an MiAP of 45.5% (XRCE), followed by an MiAP of 43.7% (LEAR). Both runs make use of multi-modal information. The best results in the evaluation per example were achieved in a visual configuration with 68% F_{eb} (ISIS) and in a multi-modal configuration with 66% OS-FCS (XRCE).

Results for the visual configuration: Table 10.7 shows the results of the best run of each group that participated in the visual configuration, evaluated with all three evaluation measures. The best results in the visual configuration were achieved by the ISIS team in terms of MiAP and F_{eb} , and the XRCE team in terms of OS-FCS. Both teams achieve close results in the concept-based evaluation (1.7% difference), while there is a larger gap with significant differences in the example-based evaluation (4.1% and 4.4%).

Table 10.7.: Summary of the VCDT 2010 results for the evaluation per concept in the visual configuration sorted by MiAP.

Group ID	Descriptor	Classifier	Rank	MiAP	Rank	F_{eb}	Rank	OS-FCS
ISIS	various SIFT	SVM	1	0.41	1	0.68	8	0.60
XRCE	colour and SIFT as Fisher vectors	linear SVM	6	0.39	6	0.64	1	0.65
LEAR	colour and SIFT as Fisher vectors	k -NN	9	0.36	15	0.58	28	0.39
HHI	opponent SIFT, sharpness feature	multi-kernel SVM	11	0.35	7	0.63	2	0.64
IJS	global and local	predictive clustering trees	15	0.33	14	0.60	10	0.60
BPACAD	HoG + GMM	linear logistic regression	20	0.28	30	0.43	27	0.44
Romania	colour histograms	LDA*, weighted average retrieval rank	21	0.26	22	0.53	15	0.56
INSUNHIT	SIFT	naïve-bayes NN	23	0.24	38	0.21	31	0.37
LSIS	extended local binary patterns	linear SVM	24	0.23	23	0.53	19	0.54
LIG	colour SIFT	non-linear SVM	30	0.23	27	0.48	20	0.53
MEIJI	SIFT	NN	31	0.22	18	0.56	34	0.36
WROCLAW	global colour and texture	NN, Penalised Discriminant Analysis	34	0.19	26	0.48	30	0.38
MLKD	C-SIFT	ensemble classifier chains	40	0.18	37	0.22	37	0.36
UPMC	global and local	Ranking SVMs	42	0.15	43	0.17	40	0.35
CEALIST	global and local	shared boosting	43	0.15	29	0.45	26	0.46

Table 10.8.: Summary of the VCDT 2010 results for the evaluation per concept in the textual configuration sorted by MiAP.

Group ID	Descriptor	Classifier	Rank	MiAP	Rank	F_{eb}	Rank	OS-FCS
MLKD	250 most frequent tags	ensemble classifier chains	1	0.23	1	0.26	1	0.37
DCU	document expansion, relationships, EXIF	XOR operation between features	2	0.23	2	0.18	2	0.30

Table 10.9.: Summary of the VCDT 2010 results for the evaluation per concept in the multi-modal configuration sorted by MiAP.

Group ID	Descriptor	Classifier	Rank	MiAP	Rank	F_{eb}	Rank	OS-FCS
XRCE	colour and SIFT as Fisher vectors, 698 most frequent tags	linear SVM	1	0.46	1	0.66	1	0.66
LEAR	colour and SIFT as Fisher vectors, 698 most frequent tags	k -NN	3	0.44	3	0.60	5	0.41
MEIJI	SIFT, tf-idf of tags	NN	6	0.33	6	0.57	3	0.43
CNRS	SIFT, tags	SVM	9	0.30	9	0.35	4	0.42
MLKD	C-SIFT, 250 most frequent tags	fusion of T and V classifier outputs	14	0.24	13	0.26	12	0.38
UPMC	global and local, Porter stemming on tags	Ranking SVMs	15	0.18	15	0.19	15	0.35

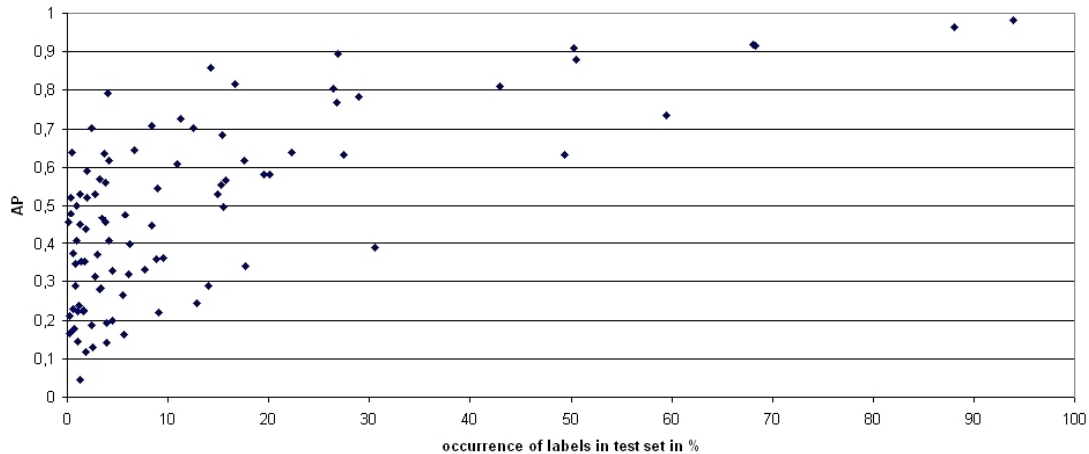


Figure 10.3.: Frequency of labels in the test set plotted against the best iAP of runs.

Results for the textual configuration: The results for the two textual runs are presented in Table 10.8. Both groups achieve similar results in the concept-based evaluation. However, the example-based evaluation shows a significant difference between the results in both measures.

Results for the multi-modal configuration: Table 10.9 depicts the results for the best multi-modal configuration of each group. As already stated, the run of XRCE achieves the best overall results in terms of MiAP and OS-FCS. In terms of OS-FCS, the results of XRCE in the multi-modal configuration are around 23% better than the second best configuration of the MEIJI team.

10.7.4. Comparison of the three configurations

The best results for each concept are summarised in Table 10.10. On average, the concepts could be detected with a iAP of 48%, considering the best results per concept from all configurations and submissions. Out of 93 concepts, 61 could be annotated most successfully with a multi-modal approach, 30 with a visual approach, and two with a textual one. Most of the concepts were classified best by one configuration of the XRCE, ISIS, or LEAR group.

The best classified concepts are those from the mutually exclusive categories: *Neutral Illumination* (98.2% iAP, 94% F), *No-Visual-Season* (96.5% iAP, 88% F), *No Persons* (91.9% iAP, 68% F), *No Blur* (91.5% iAP, 68% F). Following, the concepts *Outdoor* (90.9% iAP, 50% F), *Sky*, (89.5% iAP, 27% F), *Day* (88.1% iAP, 51% F), and *Clouds* (85.9% iAP, 14% F) were annotated with a high iAP. The concepts with the worst annotation quality were *abstract* (4.6% iAP, 1% F), *old-person* (11.6% iAP, 2% F), *work* (13.1% iAP, 3% F), *technical* (14.2% iAP, 4% F), *Graffiti* (14.5% iAP, 1% F), and *boring* (16.2% iAP, 6% F). The percentages in parentheses denote the detection performance in iAP and the frequency (F) of the concept occurrence in the test set. Although there is a trend that frequently occurring concepts can be detected better, this does not hold for all concepts. Figure 10.3 shows the frequency of concepts in the test set plotted against the highest AP achieved by any submission.

Although the performance of the textual runs is much lower on average than in the visual and multi-modal runs, there are two concepts that can be annotated most successfully in a textual configuration: *skateboard* and *abstract*. The concept *skateboard* was only annotated in six images of the test set and twelve of the training set. In the user tags of three images, the word “skateboard” was present, while two images have no user tags and the sixth image does not contain words such as “skateboard” or “skateboarding”. It seems as if there is not enough visual information

Table 10.10.: Best annotation performance per concept in the VCDT in 2010, achieved by any team in any configuration, in terms of iAP.

Concept	iAP	Team	Conf.	Concept	iAP	Team	Conf.
Partylife	0.408	LEAR	M	Food	0.635	XRCE	M
Family_Friends	0.555	ISIS	V	Vehicle	0.546	XRCE	M
Beach_Holidays	0.531	LEAR	M	Aesthetic_Impression	0.339	ISIS	V
Building_Sights	0.609	ISIS	V	Overall_Quality	0.289	ISIS	V
Snow	0.530	XRCE	M	Fancy	0.245	LEAR	M
Citylife	0.566	XRCE	M	Architecture	0.361	ISIS	V
Landscape_Nature	0.816	ISIS	V	Street	0.398	ISIS	V
Sports	0.186	ISIS	V	Church	0.288	LEAR	M
Desert	0.210	MEIJI	M	Bridge	0.224	XRCE	M
Spring	0.229	XRCE	M	Park_Garden	0.476	XRCE	M
Summer	0.332	ISIS	V	Rain	0.167	LEAR	M
Autumn	0.438	XRCE	M	Toy	0.370	XRCE	M
Winter	0.522	XRCE	M	MusicalInstrument	0.179	CNRS	M
No_Visual_Season	0.965	ISIS	V	Shadow	0.194	ISIS	V
Indoor	0.639	ISIS	V	bodypart	0.320	XRCE	M
Outdoor	0.909	XRCE	M	Travel	0.199	ISIS	V
No_Visual_Place	0.634	ISIS	V	Work	0.131	XRCE	V
Plants	0.805	ISIS	V	Birthday	0.169	LEAR	M
Flowers	0.618	XRCE	M	Visual_Arts	0.389	ISIS	V
Trees	0.702	ISIS	V	Graffiti	0.145	XRCE	M
Sky	0.895	XRCE	M	Painting	0.281	LEAR	M
Clouds	0.859	XRCE	M	artificial	0.219	LEAR	M
Water	0.725	XRCE	M	natural	0.734	LEAR	M
Lake	0.353	XRCE	M	technical	0.142	ISIS	V
River	0.351	LEAR	M	abstract	0.046	DCU	T
Sea	0.568	XRCE	M	boring	0.162	ISIS	V
Mountains	0.561	ISIS	V	cute	0.632	XRCE	M
Day	0.881	XRCE	M	dog	0.702	XRCE	M
Night	0.646	XRCE	M	cat	0.374	LEAR	M
No_Visual_Time	0.811	XRCE	M	bird	0.589	XRCE	M
Sunny	0.496	ISIS	V	horse	0.521	MEIJI	M
Sunset_Sunrise	0.791	XRCE	M	fish	0.480	MEIJI	M
Still_Life	0.445	LEAR	M	insect	0.499	XRCE	M
Macro	0.529	ISIS	V	car	0.455	XRCE	M
Portrait	0.684	XRCE	M	bicycle	0.449	XRCE	M
Overexposed	0.225	XRCE	M	ship	0.237	MEIJI	M
Underexposed	0.328	XRCE	M	train	0.347	XRCE	M
Neutral_Illumination	0.982	XRCE	M	airplane	0.640	MEIJI	M
Motion_Blur	0.284	ISIS	V	skateboard	0.455	DCU	T
Out_of_focus	0.223	ISIS	V	female	0.616	ISIS	V
Partly_Blurred	0.769	ISIS	V	male	0.782	XRCE	M
No_Blur	0.915	ISIS	V	Baby	0.407	XRCE	M
Single_Person	0.582	XRCE	M	Child	0.312	XRCE	M
Small_Group	0.359	XRCE	M	Teenager	0.266	LEAR	M
Big_Group	0.466	ISIS	V	Adult	0.582	ISIS	V
No_Persons	0.919	XRCE	M	old_person	0.116	LEAR	M
Animals	0.708	XRCE	M				

available to learn this concept, while the textual and multi-modal approaches can make use of the tags and extract the correct concept for at least half of the images.

Further, one can see a great difference in annotation quality between the old concepts from 2009 that were carefully annotated by experts (number 1-52) and the new concepts (number 53-93) annotated with the service of MTurk. The average annotation quality in terms of iAP for the old concepts is 57%, while it is 37% for the new concepts. The reason for this is unclear. One reason may lie in the quality of the annotations of the non-experts. However, recent studies have found that the quality of crowdsourced annotations is similar to the annotation quality of experts (compare Chapter 4.2). Another reason could be the choice and difficulty of the new concepts, as some of them are not as obvious and objective as the old ones. Further, some new concepts are special and their occurrence in the dataset is lower (7% in average) than the occurrence of the old concepts (17% in average). One possibility of determining the reliability of a test collection is to calculate Cronbach's alpha value, as proposed by Bodoff (2008). It defines a holistic measure of reliability and analyses the variance of individual test items and total test scores. The measure returns a value ranging between zero and one, for which larger scores indicate a higher reliability. The Cronbach's alpha values show a high reliability for the whole test collection, with 0.991 for the queries assessed by experts and 0.956 for the queries assessed by MTurk. Therefore, the scores point to a reliable test collection for both the manual expert annotations and the crowdsourced annotations and cannot explain the differences in iAP.

In the visual and multi-modal configurations, discriminative classifiers and nearest-neighbour approaches dominate. As in the previous cycles, most teams applied SIFT and colour SIFT descriptors for visual classification. Some teams additionally used global descriptors such as colour histograms. Many teams analysed the X most occurring Flickr tags and built a binary vector of the most occurring tags for each image to incorporate textual information. One group experimented with document expansion approaches.

10.8. Evolvement of concept detection performance

In the following, the focus is placed on the improvement and evolvement of concept detection performance in the three evaluation cycles.

10.8.1. Evolvement from ImageCLEF 2009 to ICPR 2010

The Photo Annotation task in ImageCLEF 2009 posed a similar problem as did the annotation challenge in the ICPR contest. In both evaluation cycles, the participants were asked to annotate a set of Flickr images with 53 visual concepts. In 2009 and 2010, the training set consisted of 5,000 images, while in 2010, an additional validation set of 3,000 images was provided. These 3,000 images belonged to the test set of 2009. The test was performed on 13,000 and 10,000 images in 2009 and 2010, respectively. Therefore, a comparison of the annotation performance on the 10,000 images of the test set used in both evaluation cycles can be made. Nevertheless, one has to keep in mind that in 2010, the algorithms could be trained with $\sim 40\%$ more training data. An increase in detection is therefore not necessarily caused by better annotation systems. Figure 10.4 illustrates the percentage of occurrence of each concept in the datasets. It can be seen that most concepts are equally distributed in the different sets, while the percentage of occurrence between concepts varies significantly.

Table 10.4 lists the annotation performance per concept in terms of EER and AUC for the evaluation cycles in 2009 and 2010. The AUC for the concepts in 2010 is greater or equal to the results in 2009. Also, the EER scores prove better or equal performance in 2010. That means that there was no decline in the annotation performance for a concept. In numbers, the concepts could

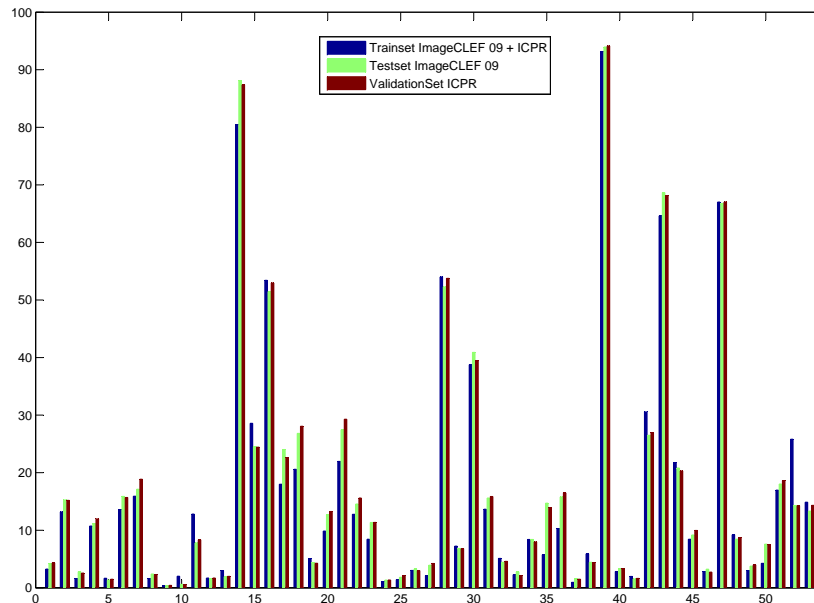


Figure 10.4.: Concept occurrences for the training set of the VCDT in 2009 and at ICPR 2010, the test set of 2009, and the validation set of ICPR 2010. The x-axis denotes the number of the concept and the y-axis shows its occurrence in percent.

be detected best by any run with an EER of 23% and an AUC of 84% in 2009. This improved to an annotation performance of 20.8% EER and 86.5% AUC per concept in 2010. However, a one-way analysis of variance reveals that the improvements per concept are not significant.

Evaluated on an example basis, the photos could be annotated correctly by 85.1%, considering the best result per photo of any run in 2010. This is a decrease by 4.5% in comparison to 2009. Also, the best run in 2010 for the example-based evaluation has a lower score (78.4%) than that in 2009 (81%). In 2010, the classification performance for each image ranged between 59% and 100%, while it ranged between 68.7% and 100% in 2009. Figure 10.5 shows two images for both evaluation cycles with the lowest detection rate in terms of OS, and the annotation rate in the other year for the same photo. It can be seen that the images that could not be annotated well in 2009 were annotated with a similar quality in 2010, while the ones that were not annotated well in 2010 were annotated much more successfully in 2009.

Summarising, the comparison shows that there was a small but not significant improvement in the annotation performance evaluated on a concept basis, while the results became slightly worse when evaluated on an example basis. Although this may seem contradictory, these results are reasonable. In contrast to the concept-based measures, the OS is directly dependent on the threshold for mapping the confidence values into a binary decision. If this threshold is not chosen carefully, the results are directly influenced, while the threshold does not affect the EER and AUC results. Further, the OS considers all labels per image. If there are a few images that could not be annotated with reasonable results, these low scores influence the average annotation behaviour. In addition, the OS penalises annotations that violate real-world knowledge provided in the ontology. Some annotation systems do not take into account concepts that condition each other through the hierarchy, or concepts that exclude each other. While the concept-based evaluation does not consider the relations between concepts, the example-based evaluation with the OS assigns violation costs in these cases. It seems as though the participants of 2009 set a higher value on these cases than in the ICPR benchmark.

Table 10.11.: Best annotation performance per concept in terms of EER and AUC at the ICPR and ImageCLEF 2010 VCDT task. Bold numbers indicate better performance.

Concept	EER ICPR	EER 2010	AUC ICPR	AUC 2010
Partylife	0.222	0.228	0.865	0.853
Family_Friends	0.216	0.222	0.858	0.853
Beach_Holidays	0.128	0.117	0.929	0.941
Building_Sights	0.178	0.181	0.903	0.898
Snow	0.191	0.141	0.892	0.929
Citylife	0.222	0.225	0.850	0.853
Landscape_Nature	0.119	0.114	0.952	0.953
Sports	0.290	0.293	0.781	0.768
Desert	0.182	0.122	0.919	0.930
Spring	0.201	0.161	0.862	0.914
Summer	0.233	0.241	0.833	0.824
Autumn	0.184	0.157	0.882	0.923
Winter	0.211	0.173	0.875	0.916
No_Visual_Season	0.248	0.257	0.822	0.820
Indoor	0.239	0.238	0.844	0.851
Outdoor	0.238	0.160	0.843	0.923
No_Visual_Place	0.270	0.253	0.811	0.830
Plants	0.181	0.181	0.903	0.904
Flowers	0.186	0.160	0.889	0.914
Trees	0.158	0.159	0.916	0.919
Sky	0.104	0.104	0.959	0.960
Clouds	0.099	0.089	0.963	0.969
Water	0.160	0.146	0.914	0.922
Lake	0.163	0.163	0.912	0.911
River	0.144	0.140	0.926	0.924
Sea	0.115	0.102	0.947	0.959
Mountains	0.117	0.115	0.948	0.950
Day	0.222	0.200	0.866	0.884
Night	0.155	0.139	0.923	0.933
No_Visual_Time	0.229	0.207	0.856	0.874
Sunny	0.267	0.273	0.812	0.808
Sunset_Sunrise	0.084	0.072	0.962	0.980
Still_Life	0.216	0.221	0.858	0.857
Macro	0.238	0.234	0.841	0.841
Portrait	0.175	0.178	0.905	0.901
Overexposed	0.244	0.213	0.832	0.851
Underexposed	0.189	0.187	0.889	0.880
Neutral_Illumination	0.260	0.258	0.811	0.818
Motion_Blur	0.277	0.292	0.785	0.785
Out_of_focus	0.179	0.220	0.905	0.846
Partly_Blurred	0.208	0.195	0.872	0.884
No_Blur	0.220	0.211	0.861	0.874
Single_Person	0.248	0.254	0.834	0.823
Small_Group	0.249	0.256	0.827	0.818
Big_Group	0.167	0.178	0.912	0.898
No_Persons	0.209	0.203	0.874	0.879
Animals	0.209	0.168	0.872	0.912
Food	0.169	0.148	0.916	0.929
Vehicle	0.226	0.212	0.855	0.864
Aesthetic_Impression	0.375	0.375	0.671	0.670
Overall_Quality	0.387	0.379	0.661	0.666
Fancy	0.421	0.419	0.614	0.614
Mean	0.208	0.199	0.866	0.873



Worst detection performance in 2009 with 68.7%.
Detection score in 2010: 68.6%



Worst detection performance in 2009 with 68.9%.
Detection score in 2010: 67.5%



Worst detection performance in 2010 with 59.0%.
Detection score in 2009: 79.1%



Worst detection performance in 2010 with 60.5%.
Detection score in 2009: 82.5%

Figure 10.5.: Images with the lowest detection rate in terms of OS in 2009 and 2010.

10.8.2. Evolvement from ICPR 2010 to ImageCLEF 2010

The VCDT at ICPR 2010 contest and in the ImageCLEF 2010 benchmark used the same training and test set. In the case of the ICPR benchmark, the training set was further divided into training and validation set, but in general, the participants were allowed to use both parts for training and learning. The major difference between both cycles lies in the extension of the number of concepts from 53 to 93, the use of Flickr tags in 2010, and the differentiation of the configuration. The following comparison concentrates on the 52 concepts that were used in both evaluation cycles.

Table 10.11 shows the results per concepts by any run in terms of EER and AUC for both cycles. On average, the annotation performance was improved from 20.8% to 19.9% EER, and from 86.6% to 87.3% AUC. For 16 concepts in terms of EER and 17 in terms of AUC, the runs from ICPR scored better. The VCDT 2010 runs achieved better scores for 32 concepts in terms of EER and 32 concepts in terms of AUC. These results are again not significant when analysed with a one-way analysis of variance. One also has to keep in mind that the submissions of the ICPR contest were optimised for the EER/AUC measures, while those from ImageCLEF 2010 were optimised for the iAP measure. A comparison of the best AP values for each concept in both cycles show that the MiAP improved from 54% in the ICPR benchmark to 57% in ImageCLEF 2010. Still, this improvement is not significant.

10.8.3. Discussion

Over all years, the best results in image annotation were obtained using discriminative classifiers. The classifier itself varied throughout the years, and the descriptors differed as well, although there

is an important trend in applying various SIFT descriptors.

In 2009, ISIS obtained the best result, using a large variety of local descriptors extracted from different interest points and grids represented in a bag-of-words-descriptor and χ^2 -SVM classifiers. For the photo-based evaluation, the XRCE group achieved the best results in 2009, using local colour and texture features, and a combination of Fisher-kernel SVMs and logistic models. In the VCDT at ICPR 2010, a spectral regression approach with various SIFT features from the CVSSPRet group scored best in the concept- and example-based evaluation. In ImageCLEF 2010, ISIS achieved the best results in the visual configuration for the MiAP and F_{eb} measures with an improved system of their ICPR 2010 submission. XRCE achieved best results in terms of OS-FCS in the visual configuration, with a linear SVM and various Fisher vector features and global colour features. In the textual configuration, MLKD scored best with an approach based on ensemble classifier chains and textual features that represent the 250 most frequent user tags. In the multi-modal configuration, XRCE scored best for all three measures. In addition to the visual features, they derive textual features from the 698 most frequent Flickr tags.

The knowledge provided in form of an ontology was not further considered by most groups. Some groups applied post-processing steps that adapt the probability of the presence of a particular concept by analyzing its likely relationships, such as the XRCE run in 2009, or the HHI and DCU runs in 2010. The usefulness of combining different modalities could be proven in ImageCLEF 2010. Overall, the multi-modal runs obtained the best scores. Additionally, all groups that participated in several configurations achieved best scores in their multi-modal runs.

10.9. Outlook

At the time of writing, the VCDT of ImageCLEF 2011 is still running. The evaluation objective lies in the automated detection of visual concepts, including sentiments, and in a concept-based retrieval scenario on 200,000 Flickr images. The participants are asked to answer 40 topics which are constructed based on query logs. The retrieval can be solved with a logical connection of concepts of the PTO. Both the sentiment annotation and the concept-based retrieval may consider visual, textual, or multi-modal information. Relevance assessment is performed with MTurk and uses pooling for the concept-based retrieval task. The evaluation considers iAP, infAP, example-based F-Measure, and Semantic R-Precision (introduced in Chapter 9).

10.10. Summary

This chapter illustrates, analyses, and discusses the evaluation activities conducted in the ImageCLEF evaluation campaign for the Photo Annotation task. The evaluation objectives, test collections, approaches of participants, results, and the evolution over three evaluation cycles are demonstrated. In these cycles, a total of 29 research groups participated and submitted 180 run configurations. Next to the achievements of the various groups, the contributions of my work on test collections and performance measures have been tested and evaluated in a real-world benchmarking scenario. Results show their applicability and the acceptance of the community on the evaluation methodology.

Part IV.

Conclusions

11. Conclusion and future work



This thesis has explored evaluation methodologies for the performance assessment of image annotation approaches. Specifically, the introduction of semantic relatedness measures into performance assessment was addressed and its consequences were thoroughly analysed. Findings were related to a user model and user requests; they were further incorporated into current benchmarking activities. In this chapter, the outcomes of the thesis and requirements on evaluation methodologies for image annotation are summarised. Section 11.1 recapitulates research performed and describes contributions. Subsequently, the focus is placed on the answering of the research questions in Section 11.2. Section 11.3 raises limitations of this work, while, finally, Section 11.4 outlines thoughts on future work.

11.1. Summary

In this thesis, evaluation methodologies for the evaluation of VIR and annotation systems have been proposed, analysed, and discussed. The work followed the hypothesis that the evaluation of image annotation approaches poses its own demands on test methodologies and evaluation design. The focus was placed to the three issues on determining visual concepts in accordance to user requirements, on obtaining low-cost relevance judgements, and on the definition of novel performance measures for the fine-grained evaluation of annotation approaches.

Part I described the context and background of this work on image annotation and evaluation in IR, and detailed the state-of-the-art in evaluation methodologies for image annotation performance assessment. Specifically, Chapter 4 gave a thorough analysis of the usage of performance measures in image annotation and presented open issues of current approaches with respect to test collections, relevance assessment, and performance measures.

Next, Part II focused on visual test collections by discussing problems of visual concept definition and relevance assessment. Chapter 5 began with the definition of appropriate visual concepts for indexing approaches. In particular, a user study was performed in order to assess which concepts users prefer to organise private photo collections. Based on these user preferences, a visual lexicon

was built and its concepts were structured in an ontology. Further, Chapter 6 explored whether the tedious process of relevance assessment in image annotation can be outsourced, and investigated the reliability of crowdsourced assessments in comparison to expert judgements. The results prove that crowdsourcing exhibits a fast and cheap possibility to obtain relevance judgements while retaining sufficient reliability. This makes crowdsourcing an interesting alternative or addition to expert annotations and pooling methods.

Part III of the thesis dealt with novel evaluation methodologies for image annotation performance assessment. In this part, I developed three novel performance measures for multi-label evaluation. They are designed for label set and ranked predictions of annotation systems, respectively, and related to a user model on concept-based image search. These measures evaluate performance based on the set of indexed concepts and consider fine-grained costs in order to determine the degree of misclassifications. Initially, in Chapter 7, semantic relatedness among concepts was estimated by measuring the path between concepts in the proposed ontology. This assumes that the number of links between two concepts is determined by their mutual similarity. However, links in an ontology do not usually provide uniform distances. For this reason, Chapter 8 investigated several semantic relatedness measures that rely on different knowledge sources, such as the WWW, Flickr, Wikipedia, or WordNet, and compared them to human judgement on semantic relatedness among concept pairs. Finally, semantic relatedness based on information from Flickr has proven to include the best statistical characteristics, while being highly correlated to human judgement at the same time. Consequently, semantic relatedness estimation based on the Flickr corpus has been included in the performance measures. Chapter 9 proposed a novel performance measure for the example-based evaluation of ranked predictions. It also makes use of the Flickr corpus to determine semantic relatedness. In contrast to the other two measures, it considers the likelihoods of the classifier predictions directly, instead of requiring a threshold which converts likelihoods into binary decisions. The problems resulting from the thresholding were demonstrated. Therefore, this measure is an addition to the other measures and either one can be applied depending on the needs of the system setup. The proposal of the novel performance measures was accompanied by a thorough evaluation of traditional performance measures, and the strengths and weaknesses of these measures were pointed out including an overall recommendation.

The outcomes of this thesis find application in the Photo Annotation task of the ImageCLEF benchmark in 2009 to 2011. Its 180 submissions constitute a representative selection of the ability of state-of-the-art image indexing systems. Chapter 10 detailed challenges, participation, conclusions, and achievements of the organisation of three cycles of the Photo Annotation task.

11.2. Contributions

The contributions of this thesis fall into three categories. They are related to user requirements and intentions, performance measures for multi-label evaluation, and visual test collections. With respect to the research questions outlined in Chapter 1.2, the contributions are summarised in the following:

- **Q1: How is image annotation evaluation performed today?**

System-based image annotation evaluation follows the Cranfield paradigm similar to other text and multimedia retrieval evaluation methodologies. Performance is assessed on a test collection, consisting of the relevance judgements, the visual concepts, and the images which are divided into training and test set. The review on evaluation in related work has shown that commonly applied image annotation performance measures can be divided into the measurement dimension, the prediction format dimension, and the relevance format dimension. The score prediction dimension is usually ignored, as nearly all measures calculate

with binary scores. In performance assessment, the evaluation measures are defined before the evaluation starts and rely to a great extent on concept-based performance measures using label set or ranked predictions. A clear dominance of concept-based Precision and Recall, followed by MAP, could be proven. Benchmarks for visual annotation approaches make use of fully assessed test collections. Only the TRECVID high-level feature extraction task includes pooling in relevance assessment.

- **Q2: Which specific needs should be addressed in the evaluation of image annotation?**

Image annotation evaluation should focus on the information needs of the user. In system-based evaluation, the user's information need is incorporated into the evaluation process by the selection of adequate concepts and the choice of appropriate evaluation measures with respect to the search scenario. The latter point is determined by the user model, and will be summarised in the answer to question Q3. The first issue is of more general nature and depends on the common way of managing private photo collections, as well as on the concepts and categories that users request. The results of the study on photo collection organisation revealed that roughly 70% of the attributes refer to content-based characteristics, 13% can be classified as representational characteristics, and 13% are related to affective properties. In particular, concepts related to the concept `people` are important and should be considered in the analysis. The review of related work on organising photo collections is in line with these results and stresses the need for indexing quality related concepts in order to give the user fast rejection possibilities of poor images. The quality concepts are not restricted to technical aspects of quality, but also include subjective opinions such as boring photos. Moreover, events, objects, and scenes play a major role. These requirements found application in the visual concept lexicon and the PTO of the ImageCLEF VCDT test collection. In contrast to related test collections, it contains a holistic view on image indexing and includes representational concepts, emotional concepts, events, and seasons, besides the usual object- and scene-related attributes.

- **Q3: What does a user model for concept-based image retrieval look like?**

The user model for concept-based image retrieval considers three requirements. First, image annotation approaches should be evaluated based on the complete set of concepts that they predict in comparison to the complete set of concepts in the ground truth. This directly relates to the search behaviour of the user. If users ask for several visual concepts to be depicted simultaneously in an image, they expect that all concepts are visible in the images of the retrieved result set. Therefore, the evaluation should judge quality in the same manner as the user would judge his satisfaction with the result set. Second, the evaluation should measure performance in a fine-grained fashion in order to distinguish badly and moderately well annotated images. The distinction into semantically related and unrelated predictions with respect to the ground truth helps to find images that are closer to the user's information need. Especially in small databases or for special topics, this reveals valuable answers for the user. Third, the performance assessment should consider the predictions in the format which is used in later processing steps of the system. This requirement addresses the fact that image annotation often denotes only one step in the concept-based retrieval process. Performance assessment based on binary predictions distort the information on retrieval performance when likelihoods are used in the later ranking of possible retrieval candidates.

- **Q4: How can the effort in relevance assessment be reduced?**

Usually, the relevance assessment is the most time- and cost-consuming part of the performance assessment. As a result, research groups focus on alternatives to the acquisition of a complete set of expert judgements. Methods include pooling, statistical sampling, and

research on minimal test collections. Recently, the acquisition of relevance judgements from non-experts with crowdsourcing approaches has been investigated. Crowdsourcing makes use of the fact that humans can perform tasks such as image annotation without special knowledge. The results of my study on the reliability of crowdsourced ground truth proves crowdsourcing as a cheap and fast way to obtain relevance for visual test collections. The majority vote of several non-expert judgements is sufficient to filter noise from the answers and generate a ground truth which is of comparable quality to that of experts. This makes crowdsourcing an interesting alternative or addition to expert annotations and pooling methods. The utilisation of user-generated tags for training corpora comprises another alternative to expert and crowdsourced judgements. The combined analysis of tags and visual image characteristics allows filtering relevant images with respect to a visual concept from irrelevant ones out of Web corpora and photo sharing sites. While this method does not deliver clean image sets, it allows for the automated download of large-scale training sets. The image sets might be cleansed in a second step.

- **Q5: Is it possible and beneficial to include semantics in the evaluation process?**

Images depict several visual concepts simultaneously. Therefore, these concepts share a semantic relationship. This thesis has worked on the hypothesis that pairs of concepts from the predicted and ground truth set follow different degrees of semantic relatedness, and that this relation should be considered in the performance assessment. With the introduction of the OS, its extension OS-FTS, and the SR-Precision, this work proposed and explored evaluation measures that incorporate semantic relatedness into the performance assessment. The measures consider the two sides of the prediction format dimension, and different corpora for estimating semantic relatedness were analysed. The OS and OS-FTS are applicable to label set predictions. They were compared to common concept-based and example-based performance measures in multi-label image annotation which are applied to label set and ranked predictions. Results show that the OS is mostly able to differentiate between random runs and real systems, and the influence on the results of predictions which are close to the correct LD of the ground truth is rather small. However, the OS favours systems that follow all ontology rules, but are not ranked at top with traditional measures. The SR-Precision measure considers ranked predictions in order to leave the thresholding problem out of the performance assessment. It is compared to other example-based measures for ranked predictions. While it exhibits a high correlation to R-Precision, its advantages are apparent in the analysis of single images. In contrast to R-Precision, the SR-Precision is able to differentiate the annotation performance of images that are assigned the same number of misclassified concepts. Results of a user experiment demonstrate that this difference in quality is also perceived and differentiated by human subjects. At the same time, the SR-Precision benefits from equal positive statistical characteristics in terms of ranking performance and stability to noise as the R-Precision. It therefore inherits an equal ability to discriminate well among the varying prediction quality of systems as well as among random annotations.

- **Q6: Which information source provides the best estimate on semantic relatedness according to desired evaluation characteristics and human judgement?**

Literature on semantic relatedness estimation classifies methods into thesaurus-based, Wikipedia-based, and Web-based approaches. Web-based approaches might be further differentiated into document-based measure and image-based measures, depending on the Web corpus used. All approaches propose an automated method to derive semantic relatedness of concept pairs. In extensive ranking and stability experiments, as well as regarding the

closeness to human semantic relatedness judgements, 14 relatedness measures of all categories have been assessed. Results for the statistical experiments reveal that four of the WordNet measures (HSO, JCN, PATH, VEC) and the Flickr-based distributional measures (FCS, FTS) exhibit a correlation to the baseline measure while retaining sufficient stability. The distributional metrics based on Yahoo and Google show different results. While results with Yahoo are stable, evaluation using the metric based on the Google search engine is sensitive to noise. Besides, it is essential that the semantic relatedness metrics cover the semantic relatedness of concepts as closely as possible to human opinion. A new benchmark for semantic relatedness of visual concepts is proposed, as the state-of-the-art only focuses on the measurement of semantic relatedness for word pairs. Results show that the Flickr measures inhibit the highest semantic relatedness to human opinion. The WordNet measures do not perform well on the visual concept dataset. In line with the statistical results, JCN and PATH estimate semantic relatedness closest to human judgements of all WordNet measures. These results are contrary to the semantic relatedness judgement on the Gracia and Mena (2008) dataset of word pairs. Further, relatedness estimation using Yahoo inhibits a significantly higher correlation to human judgement than with Google.

To summarise, outcomes of this thesis result in a user model for concept-based image retrieval, a fully assessed image annotation test collection, and a number of novel performance measures for image annotation evaluation.

11.3. Limitations

Besides the contributions of this thesis with respect to performance measures, test collections, and user preferences, some issues are not yet solved. These limitations are discussed in the following and solutions are proposed. They relate to problems of performance measures for image annotation, issues on evolving corpora for semantic relatedness estimation, the comparability of classifier score predictions, system-based meta-evaluation of performance measures, the lack of an extensive user study to verify the proposed user model, and to spam in crowdsourcing judgements.

- **Limitations of performance measures for image annotation**

Several limitations of common performance measures for image annotation were resolved in this thesis. The basic idea to introduce semantics in the evaluation process was tackled with the OS measure and the determination of semantic relatedness by the hierarchy and relations of an ontology. Its limitation, the arbitrary structure of the ontology, was verified with user judgements on semantic relatedness and compared to several alternatives in order to estimate semantic relatedness. These studies resulted in the improvement of the OS with semantic relatedness estimation from Flickr. Both of these measures rely on label set predictions. This potential limitation, depending on the system setup, was addressed by the introduction of the SR-Precision which works for ranked predictions. However, all three measures require a fully assessed test collection to be applicable. It has not been investigated how methods such as pooling can be efficiently adapted to example-based annotation evaluation. In the case of concept-based evaluation, the assumption of pooling is that all relevant images for a concept will be ranked at the top X ranks in at least one run and therefore contribute to the pool. This assumption works as long as the concepts are not too general. It does not hold, for example, for concepts like *Outdoor*, which is depicted in about 54% of the images in the VCDT test collection if X is chosen in the usual range. In example-based evaluation, an image has to be assessed with all concepts that the test collection supports. The straightforward solution is to assess all concepts for

each image that occurs in the first X ranks of any result list for any concept. Intuitively, the author doubts whether this will reduce the amount of relevance judgements needed. Further research has to be performed to make example-based measures scalable to large test collections with incomplete relevance judgements.

- **Evolving corpora for semantic relatedness estimation**

Results of the semantic relatedness experiments in Chapter 8 show that the Flickr-based relatedness measures FTS and FCS imply the best statistical characteristics and are highly related to human judgement at the same time. For these reasons, the FTS was recommended as the best measure to estimate semantic relatedness for visual concepts. However, Flickr hosts a corpus of images and user tags which is constantly growing and changing. The FTS measure considers the co-occurrence of two concepts and relates it to the total number of appearances of each concept in order to determine the relatedness value. The frequencies obtained at one point in time might differ from concept frequencies taken at a different point in time. Experiments in Chapter 9 consider the FTS in the evaluation of the runs of the Photo Annotation task from 2009 and 2010. The relatedness scores have been determined at different points in time from Flickr. However, ranking and stability results do not show any differences between the years. Therefore, it seems as though the general tendencies on tag distributions hold over time.

In general, the evolving corpus is no problem for performance assessment in benchmarking scenarios, as all runs are evaluated at a specific date. However, it might be beneficial to save snapshots of costmaps at the point of evaluation and make them available to the research community. They could be distributed as one part of the test collection. This assures that the performance measurement is conducted with equal premises when evaluating runs on the same collection after the official performance assessment. As an alternative, ontologies could be used with the hierarchical distance function as proposed in Chapter 7. Results on the correlation experiment to human judgements have demonstrated that the PTO shows a good estimate on semantic relatedness. However, the ontology has to be adapted for any new concept. As this changes the structuring of concepts in the hierarchy, the relatedness experiment needs to be repeated. Results have demonstrated that the OS relying on the first version of the PTO fits well to human opinion, but there is no guarantee that this also holds for later versions. On the contrary, experiences with large ontologies such as WordNet show that relatedness estimation between visual concepts was not convincing for nine different measures.

- **Comparability of classifier score predictions**

For the performance assessment of ranked predictions, the example-based and concept-based evaluation makes the assumption that likelihoods are comparable across concept categories and documents, respectively. While the concept-based evaluation is widely established and predictions throughout several documents seem to be comparable, there is no experiment which proves that this is also the case for likelihoods across different concepts. Possibly, likelihoods have to be adapted to be comparable among different concepts depending on the classifier used. Especially, most state-of-the-art image annotation systems transform the multi-label classification problem into several single-label classification problems, or consider a multi-class categorisation approach for mutually exclusive concepts and a binary classification for optional concepts, as could be seen in the submissions to the Photo Annotation task. However, to the knowledge of the author, there is no study which analyses the ranges of likelihood values of different classifiers for different concepts. As the example-based performance measures for ranked predictions heavily rely on an ordering of the concept likelihoods,

this issue directly affects the outcome of the performance assessment. A detailed study of this problem remains an issue of future work.

- **Issues on system-based meta-evaluation**

The system-based comparison of the effectiveness of performance measures is difficult. As illustrated in this thesis, empirical studies that investigate ranking and stability behaviour, or an analysis of mathematical properties of measures can be conducted. The results of the stability experiments, especially in Chapter 9, demonstrated that all performance measures for ranked predictions enclose a high stability. The discrimination abilities among the measures was low. Questions of why the performance measures act this insensitively to noise and of whether the evaluation setup is adequate to evaluate stability remain. In the case of the correlation studies, experiments need a baseline to which measures can be compared, or results can only determine relative characteristics of measures to each other. In user-centred evaluation this baseline is defined with respect to user requirements and user judgements, but it would actually require a “meta-metric” to perform correlation experiments solely system-based; this is a circular reference problem. In the experiments on semantic relatedness metrics in Chapter 8, the example-based F-measure was chosen as baseline performance measure due to its superior characteristics on random numbers and overannotation, analysed in Chapter 7. However, the comparison of fine-grained performance assessment with binary performance assessment is not optimal as these measures are not intended to behave equally. The behaviour of measures should better be compared to human preferences on evaluation measures. This has been analysed in this thesis for the ability of metrics to estimate semantic relatedness according to human opinion. However, the ranking and stability preferences of performance measures judged by users should additionally be taken into account. Such a large-scale user study was out of scope of this thesis and remains an issue of future work.

- **User study to verify the user model on concept-based search**

In the thesis, a small user experiment was conducted in order to validate the proposed user model on concept-based search. The results are promising and demonstrate a benefit in applying a fine-grained evaluation based on semantic relatedness. Users are able to distinguish among different qualities of concept sets in cases for which traditional performance measures assign the same score to both sets. While these results point to an acceptance of the user model on concept-based search, the user model has to be confirmed in a large-scale user experiment. Final conclusions on the benefits of semantics in evaluation assessment can only be drawn when users obtain the possibility of interacting with concept-based retrieval systems in a real-world scenario. The MTurk experiment has shown the difficulties in designing an experiment which verifies the requirements of the user model without posing leading questions. Probably, a “traditional” user study in which the human subjects can be monitored and which allows for an interaction with subjects denotes a better validation scenario. The user experiments on performance measures in this thesis only constitute a first step in this direction.

- **Spam in crowdsourcing judgements**

The influence of variations in relevance assessment on system ranking performance is low. Results of the study in Chapter 6 prove that variations in relevance do not significantly affect system ranking and that averaged non-expert judgements retain the possibility of discriminating among systems. These results are in line with studies on the influence of variations in relevance in text retrieval (Voorhees (2000)). The stability experiments prove this in a similar way. The introduction of noise in the ground truth of up to 10% can be

regarded as variation of relevance assessment, or as spam as it might occur in crowdsourcing experiments. In many cases, the noise had a low influence on the evaluation results and performance assessment produced stable rankings, especially for performance measures that consider ranked predictions. The influence of variations in relevance - or spam judgements - on ranking behaviour is crucial for crowdsourcing experiments. The crowdsourcing study on the reliability of relevance judgements of non-experts from November 2009 was conducted quite early with respect to other crowdsourcing experiments in IR. In the last two years, research experiments on platforms such as MTurk became increasingly popular and many research groups use the results in their work. At the same time, the population of workers started to shift from people who crowdsourced for pleasure to people that need crowdsourcing to earn money and that are therefore potentially interested in maximizing money with low effort (Ross et al. (2010)). This might be one reason why the number of spam answers are starting to increase. It has to be investigated if the simple mechanism of a majority vote, as proposed in this work, is still sufficient to gather a reliable ground truth in today's crowdsourcing environment. Probably, gold standard questions, qualification tests, and other mechanisms to ensure quality should be integrated. Experiments on MTurk for the relevance assessment in ImageCLEF 2011, which are running at the time of writing, suggest that these steps are necessary to assure quality. However, the benefits of crowdsourcing in terms of a cheap and fast assessment remain, while posing new challenges for automated spam detection at the same time.

11.4. Directions on future work

Future work will consider the points on incomplete relevance in example-based annotation evaluation, the verification of whether classifier prediction scores are comparable across concepts, user studies to determine preferences in ranking behaviour and to verify the user model, and methods for spam detection in crowdsourcing experiments, as discussed above. Besides, the relevance format dimension of performance measures has not yet been considered for example-based annotation evaluation. Reasons lie mainly in the unavailability of test collections comprising graded-relevance judgements. However, graded relevance determines another way of integrating semantics in the evaluation process. In this case, the relation among different votes of assessors is modelled. This means allows introducing subjectivity in assessment into the evaluation process. A similar idea was followed with the annotator agreements that were incorporated in the OS in Chapter 7. However, further experiments showed no difference in system ranking for performance assessment with and without the use of annotator agreements. Therefore, this idea was not further followed in this thesis. The introduction of the subjectivity of relevance judgements might lead to more interesting insights. Further, the generalisation of the proposed performance measures with respect to the evaluation of video annotation, music annotation, or text annotation poses interesting questions. Which corpora can be applied best for semantic relatedness estimation among concepts in music? Can the OS-FTS and SR-Precision measures be directly applied to video annotation evaluation? In how far do the results differ from the results obtained in image annotation evaluation?

Part V.

Appendix

A. List of visual concepts in ImageCLEF

A.1. Visual concepts in ImageCLEF 2009 and ICPR 2010

Table A.1.: Visual concepts used in ImageCLEF 2009 and at ICPR 2010 with the corresponding categories in the Photo Tagging Ontology.

#	Concept	Category in ontology
0	Partylife	SceneDescription.AbstractCategories.Partylife
1	Family_Friends	SceneDescription.AbstractCategories.FamilyFriends
2	Beach_Holidays	SceneDescription.AbstractCategories.BeachHolidays
3	Building_Sights	SceneDescription.AbstractCategories.BuildingsSights
4	Snow	SceneDescription.AbstractCategories.SnowSkiing
5	Citylife	SceneDescription.AbstractCategories.Citylife
6	Landscape_Nature	SceneDescription.AbstractCategories.LandscapeNature
7	Sports	SceneDescription.Activity.Sports
8	Desert	SceneDescription.AbstractCategories.Desert
9	Spring	SceneDescription.Seasons.Spring
10	Summer	SceneDescription.Seasons.Summer
11	Autumn	SceneDescription.Seasons.Autumn
12	Winter	SceneDescription.Seasons.Winter
13	No_Visual_Season	SceneDescription.Seasons.NoVisualCue
14	Indoor	SceneDescription.Place.Indoor
15	Outdoor	SceneDescription.Place.Outdoor
16	No_Visual_Place	SceneDescription.Place.NoVisualCue
17	Plants	LandscapeElements.Plants
18	Flowers	LandscapeElements.Plants.Flowers
19	Trees	LandscapeElements.Plants.Trees
20	Sky	LandscapeElements.Sky
21	Clouds	LandscapeElements.Sky.Clouds
22	Water	LandscapeElements.Water
23	Lake	LandscapeElements.Water.Lake
24	River	LandscapeElements.Water.River
25	Sea	LandscapeElements.Water.Sea
26	Mountains	LandscapeElements.Mountains
27	Day	SceneDescription.TimeOfDay.Day
28	Night	SceneDescription.TimeOfDay.Night
29	No_Visual_Time	SceneDescription.TimeOfDay.NoVisualCue
30	Sunny	SceneDescription.TimeOfDay.Sunny
31	Sunset_Sunrise	SceneDescription.TimeOfDay.SunsetOrSunrise
32	Canvas	Representation.Canvas
33	Still_Life	Representation.StillLife
34	Macro	Representation.MacroImage
35	Portrait	Representation.Portrait
36	Overexposed	Representation.Illumination.Overexposed
37	Underexposed	Representation.Illumination.Underexposed
38	Neutral_Illumination	Representation.Illumination.Neutral
39	Motion_Blur	Quality.Blurring.MotionBlur
40	Out_of_focus	Quality.Blurring.OutOfFocus
41	Partly_Blurred	Quality.Blurring.PartlyBlurred
42	No_Blur	Quality.Blurring.NoBlurDetectable
43	Single_Person	PicturedObjects.Persons.Single
44	Small_Group	PicturedObjects.Persons.SmallGroup
45	Big_Group	PicturedObjects.Persons.BigGroup
46	No_Persons	PicturedObjects.Persons.NoPersons
47	Animals	PicturedObjects.Animals
48	Food	PicturedObjects.Food
49	Vehicle	PicturedObjects.Vehicles
50	Aesthetic_Impression	Quality.Aesthetics.AestheticImpression
51	Overall_Quality	Quality.Aesthetics.HighGradeOverallQuality
52	Fancy	Quality.Aesthetics.Fancy

A.2. Visual concepts in ImageCLEF 2010

Table A.2.: Visual concepts used in ImageCLEF 2010 with the corresponding categories in the Photo Tagging Ontology.

#	Concept	Category in ontology	#	Concept	Category in ontology
0	Partylife	SceneDescription.AbstractCategories.Partylife	65	Visual Arts	Representation.Art
1	Family_Friends	SceneDescription.AbstractCategories.FamilyFriends	66	Graffiti	Representation.Art.Graffiti
2	Beach_Holidays	SceneDescription.AbstractCategories.BeachHolidays	67	Painting	Representation.Art.Painting
3	Building_Sights	SceneDescription.AbstractCategories.BuildingsSights	68	artificial	Representation.Impression.Artificial
4	Snow	SceneDescription.AbstractCategories.SnowSkiing	69	natural	Representation.Impression.Natural
5	Citylife	SceneDescription.AbstractCategories.Citylife	70	technical	Representation.Impression.Technical
6	Landscape_Nature	SceneDescription.AbstractCategories.LandscapeNature	71	abstract	Representation.Impression.Abstract
7	Sports	SceneDescription.Activity.Sports	72	boring	Impression.Boring
8	Desert	SceneDescription.AbstractCategories.Desert	73	cute	Impression.Cute
9	Spring	SceneDescription.Seasons.Spring	74	dog	PicturedObjects.Animals.Dog
10	Summer	SceneDescription.Seasons.Summer	75	cat	PicturedObjects.Animals.Cat
11	Autumn	SceneDescription.Seasons.Autumn	76	bird	PicturedObjects.Animals.Bird
12	Winter	SceneDescription.Seasons.Winter	77	horse	PicturedObjects.Animals.Horse
13	No_Visual_Season	SceneDescription.Seasons.NoVisualCue	78	fish	PicturedObjects.Animals.Fish
14	Indoor	SceneDescription.Place.Indoor	79	insect	PicturedObjects.Animals.Insect
15	Outdoor	SceneDescription.Place.Outdoor	80	car	PicturedObjects.Vehicles.Car
16	No_Visual_Place	SceneDescription.Place.NoVisualCue	81	bicycle	PicturedObjects.Vehicles.Bike
17	Plants	LandscapeElements.Plants	82	ship	PicturedObjects.Vehicles.Ship
18	Flowers	LandscapeElements.Plants.Flowers	83	train	PicturedObjects.Vehicles.Train
19	Trees	LandscapeElements.Plants.Trees	84	airplane	PicturedObjects.Vehicles.Airplane
20	Sky	LandscapeElements.Sky	85	skateboard	PicturedObjects.Vehicles.Skateboard
21	Clouds	LandscapeElements.Sky.Clouds	86	female	PicturedObjects.Persons.Gender.Female
22	Water	LandscapeElements.Water	87	male	PicturedObjects.Persons.Gender.Male
23	Lake	LandscapeElements.Water.Lake	88	Baby	PicturedObjects.Persons.Age.Baby
24	River	LandscapeElements.Water.River	89	Child	PicturedObjects.Persons.Age.Child
25	Sea	LandscapeElements.Water.Sea	90	Teenager	PicturedObjects.Persons.Age.Teenager
26	Mountains	LandscapeElements.Mountains	91	Adult	PicturedObjects.Persons.Age.Adult
27	Day	SceneDescription.TimeOfDay.Day	92	old_person	PicturedObjects.Persons.Age.OldPerson
28	Night	SceneDescription.TimeOfDay.Night			
29	No_Visual_Time	SceneDescription.TimeOfDay.NoVisualCue			
30	Sunny	SceneDescription.TimeOfDay.Sunny			
31	Sunset_Sunrise	SceneDescription.TimeOfDay.SunsetOrSunrise			
32	Still_Life	Representation.StillLife			
33	Macro	Representation.MacroImage			
34	Portrait	Representation.Portrait			
35	Overexposed	Representation.Illumination.Overexposed			
36	Underexposed	Representation.Illumination.Underexposed			
37	Neutral_Illumination	Representation.Illumination.Neutral			
38	Motion_Blur	Quality.Blurring.MotionBlur			
39	Out_of_focus	Quality.Blurring.OutOfFocus			
40	Partly_Blurred	Quality.Blurring.PartlyBlurred			
41	No_Blur	Quality.Blurring.NoBlurDetectable			
42	Single_Person	PicturedObjects.Persons.Number.Single			
43	Small_Group	PicturedObjects.Persons.Number.SmallGroup			
44	Big_Group	PicturedObjects.Persons.Number.BigGroup			
45	No_Persons	PicturedObjects.Persons.Number.NoPersons			
46	Animals	PicturedObjects.Animals			
47	Food	PicturedObjects.Food			
48	Vehicle	PicturedObjects.Vehicles			
49	Aesthetic_Impression	Quality.Aesthetics.AestheticImpression			
50	Overall_Quality	Quality.Aesthetics.HighGradeOverallQuality			
51	Fancy	Impression.Fancy			
52	Architecture	SceneDescription.AbstractCategories.Architecture			
53	Street	UrbanElements.Street			
54	Church	UrbanElements.Church			
55	Bridge	UrbanElements.Bridge			
56	Park_Garden	LandscapeElements.Plants.Garden			
57	Rain	LandscapeElements.Water.Rain			
58	Toy	PicturedObjects.Toy			
59	MusicalInstrument	PicturedObjects.MusicalInstrument			
60	Shadow	Representation.Illumination.Shadow			
61	bodypart	PicturedObjects.Persons.BodyPart			
62	Travel	SceneDescription.Events.Travel			
63	Work	SceneDescription.Events.Work			
64	Birthday	SceneDescription.Events.Birthday			

A.3. Visual concepts in ImageCLEF 2011

Table A.3.: Visual concepts used in ImageCLEF 2011 with the corresponding categories in the Photo Tagging Ontology.

#	Concept	Category in ontology	#	Concept	Category in ontology
0	Partylife	SceneDescription.AbstractCategories.Partylife	65	artificial	Representation.Impression.Artificial
1	Family_Friends	SceneDescription.AbstractCategories.FamilyFriends	66	natural	Representation.Impression.Natural
2	Beach_Holidays	SceneDescription.AbstractCategories.BeachHolidays	67	technical	Representation.Impression.Technical
3	Building_Sights	SceneDescription.AbstractCategories.BuildingsSights	68	abstract	Representation.Impression.Abstract
4	Snow	SceneDescription.AbstractCategories.SnowSkiing	69	boring	Impression.Felt.Boring
5	Citylife	SceneDescription.AbstractCategories.Citylife	70	cute	Impression.Felt.Cute
6	Landscape_Nature	SceneDescription.AbstractCategories.LandscapeNature	71	dog	PicturedObjects.Animals.Dog
7	Sports	SceneDescription.Activity.Sports	72	cat	PicturedObjects.Animals.Cat
8	Desert	SceneDescription.AbstractCategories.Desert	73	bird	PicturedObjects.Animals.Bird
9	Spring	SceneDescription.Seasons.Spring	74	horse	PicturedObjects.Animals.Horse
10	Summer	SceneDescription.Seasons.Summer	75	fish	PicturedObjects.Animals.Fish
11	Autumn	SceneDescription.Seasons.Autumn	76	insect	PicturedObjects.Animals.Insect
12	Winter	SceneDescription.Seasons.Winter	77	car	PicturedObjects.Vehicles.Car
13	Indoor	SceneDescription.Place.Indoor	78	bicycle	PicturedObjects.Vehicles.Bike
14	Outdoor	SceneDescription.Place.Outdoor	79	ship	PicturedObjects.Vehicles.Ship
15	Plants	LandscapeElements.Plants	80	train	PicturedObjects.Vehicles.Train
16	Flowers	LandscapeElements.Plants.Flowers	81	airplane	PicturedObjects.Vehicles.Airplane
17	Trees	LandscapeElements.Plants.Trees	82	skateboard	PicturedObjects.Vehicles.Skateboard
18	Sky	LandscapeElements.Sky	83	female	PicturedObjects.Persons.Gender.Female
19	Clouds	LandscapeElements.Sky.Clouds	84	male	PicturedObjects.Persons.Gender.Male
20	Water	LandscapeElements.Water	85	Baby	PicturedObjects.Persons.Age.Baby
21	Lake	LandscapeElements.Water.Lake	86	Child	PicturedObjects.Persons.Age.Child
22	River	LandscapeElements.Water.River	87	Teenager	PicturedObjects.Persons.Age.Teenager
23	Sea	LandscapeElements.Water.Sea	88	Adult	PicturedObjects.Persons.Age.Adult
24	Mountains	LandscapeElements.Mountains	89	old_person	PicturedObjects.Persons.Age.OldPerson
25	Day	SceneDescription.TimeOfDay.Day	90	happy	Impression.Expressed.Happy
26	Night	SceneDescription.TimeOfDay.Night	91	funny	Impression.Felt.Funny
27	Sunny	SceneDescription.TimeOfDay.Sunny	92	euphoric	Impression.Expressed.Euphoric
28	Sunset_Sunrise	SceneDescription.TimeOfDay.SunsetOrSunrise	93	active	Impression.Expressed.Active
29	Still_Life	Representation.StillLife	94	scary	Impression.Expressed.Scary
30	Macro	Representation.MacroImage	95	unpleasant	Impression.Expressed.Unpleasant
31	Portrait	Representation.Portrait	96	melancholic	Impression.Expressed.Melancholic
32	Overexposed	Representation.Illumination.Overexposed	97	inactive	Impression.Expressed.Inactive
33	Underexposed	Representation.Illumination.Underexposed	98	calm	Impression.Expressed.Calm
34	Neutral_Illumination	Representation.Illumination.Neutral			
35	Motion_Blur	Quality.Blurring.MotionBlur			
36	Out_of_focus	Quality.Blurring.OutOfFocus			
37	Partly_Blurred	Quality.Blurring.PartlyBlurred			
38	No_Blur	Quality.Blurring.NoBlurDetectable			
39	Single_Person	PicturedObjects.Persons.Number.Single			
40	Small_Group	PicturedObjects.Persons.Number.SmallGroup			
41	Big_Group	PicturedObjects.Persons.Number.BigGroup			
42	No_Persons	PicturedObjects.Persons.Number.NoPersons			
43	Animals	PicturedObjects.Animals			
44	Food	PicturedObjects.Food			
45	Vehicle	PicturedObjects.Vehicles			
46	Aesthetic_Impression	Quality.Aesthetics.AestheticImpression			
47	Overall_Quality	Quality.Aesthetics.HighGradeOverallQuality			
48	Fancy	Impression.Fancy			
49	Architecture	SceneDescription.AbstractCategories.Architecture			
50	Street	UrbanElements.Street			
51	Church	UrbanElements.Church			
52	Bridge	UrbanElements.Bridge			
53	Park_Garden	LandscapeElements.Plants.Garden			
54	Rain	LandscapeElements.Water.Rain			
55	Toy	PicturedObjects.Toy			
56	MusicalInstrument	PicturedObjects.MusicalInstrument			
57	Shadow	Representation.Illumination.Shadow			
58	bodypart	PicturedObjects.Persons.BodyPart			
59	Travel	SceneDescription.Events.Travel			
60	Work	SceneDescription.Events.Work			
61	Birthday	SceneDescription.Events.Birthday			
62	Visual_Arts	Representation.Art			
63	Graffiti	Representation.Art.Graffiti			
64	Painting	Representation.Art.Painting			

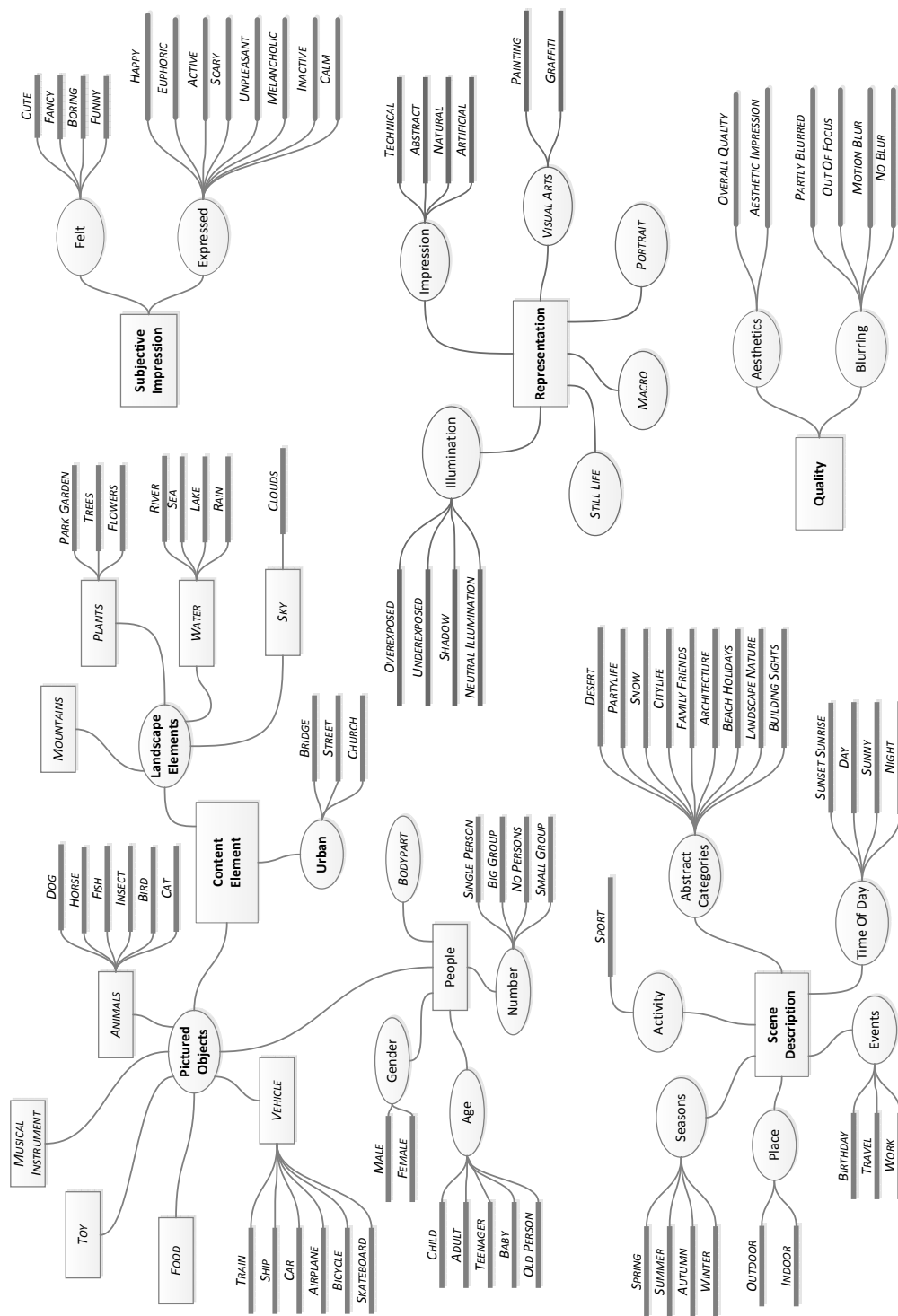


Figure A.1.: This map illustrates the concepts of the PTO of ImageCLEF 2011. See also Nagel (2011)

B. Theses

1. System-based image annotation evaluation follows the Cranfield paradigm. Performance is assessed on a test collection, consisting of the relevance judgements, the visual concepts, and the images divided into training and test set.
2. Common Information Retrieval performance measures can be classified into three dimensions. They consider the measurement direction, the prediction format dimension, and the relevance format dimension. The score prediction dimension which determines the costs that are assigned in the case of misclassifications is widely ignored. Usually, binary costs are applied.
3. The evaluation and comparison of performance measures (meta-evaluation) can be conducted in an automated manner by applying rank correlation and stability experiments. Moreover, results on random annotations and overannotation should discriminate well from the results of annotation systems. However, these statistical results only indirectly capture the user's intention on performance measure characteristics.
4. Users organise their photo collections with respect to content-based attributes, representational characteristics, and affective and emotional properties. Especially, person related attributes, events, and quality aspects of images are important. This allows users to group photos according to events or persons, and to efficiently discard low quality or boring photos.
5. Disregarding the user's information need, common visual test collections and annotation approaches mainly consider content-based attributes, while ignoring representational and affective characteristics. Visual annotation approaches have to go beyond the prediction of objective content properties and index more subjective characteristics. The ImageCLEF VCDT test collection and the associated Photo Annotation task provide a first step in this direction.
6. The user model on concept-based image retrieval incorporates three requirements on performance assessment of annotation approaches. Evaluation should assess performance based on the set of predicted concepts per image (measurement direction), it should incorporate fine-grained costs for misclassifications (score prediction dimension), and it should consider predictions in the format that is used in later processing steps in order to exclude the thresholding problem from the performance assessment (prediction format dimension).
7. Relevance assessment using crowdsourcing provides a fast and cheap alternative to expert judgements while retaining sufficient reliability for visual annotation tasks.
8. The relevance assessment process is inherently subjective. Different means of measuring the inter-annotator agreement, such as kappa statistics or accuracy, allow to determine the subjectivity of a topic and to decide on the number of assessors needed.

9. Semantic relatedness between visual concepts can be estimated by calculating co-occurrences of concepts in a corpus. Especially, the Flickr corpus as well as the hierarchical distance in the Photo Tagging Ontology provide a reliable estimate on semantic relatedness of visual concepts with respect to human judgements. In contrast, the WordNet and Wikipedia corpora are not highly correlated to human opinion.
10. Different Web search engines provide a different estimate on semantic relatedness between visual concepts. Results vary with respect to stability and ranking correlation characteristics as well as with respect to human judgements. Thus, results on the Web corpus searched with Yahoo! provide a reliable estimate on semantic relatedness on visual concepts, while results from the search with Google are sensitive to noise and inhibit only a low correlation to human judgement.
11. The comparison of image annotation performance is often difficult, as approaches are evaluated on different test collections and with different performance measures. Therefore, evaluation initiatives define challenging tasks with the goal of objectively measuring the performance of algorithms and of establishing a baseline for comparing systems. They advance the research field by providing standardised test collections and a defined evaluation methodology, while performing the evaluation in an unbiased manner.
12. Results of four cycles on the Photo Annotation task in ImageCLEF reveal that state-of-the-art systems consider discriminative classification approaches with local SIFT features or a combination of local and global features.

Acronyms

AP Average Precision. 32, 34, 46, 47, 57, 153, 160

API Application Programming Interface. 41

AUC Area-Under-Curve. 2, 32, 46, 47, 85, 86, 89, 91, 92, 107, 108, 111–116, 152, 153, 155–158, 162–165

BEP break even point precision. 33

BoW bag-of-visual-words. 17, 155

CBIR content-based image retrieval. 13, 14

CLEF Cross Language Evaluation Forum. 36, 148

corr-LDA correspondence Latent Dirichlet Allocation. 21

CRF Conditional Random Field. 18, 22

DDMC depth-dependent distance-based misclassification costs. 48, 102, 119

DIMC depth-independent distance-based misclassification costs. 48, 55

EER Equal Error Rate. 2, 32, 46, 47, 85, 86, 88, 89, 91, 92, 107, 108, 111–116, 152, 153, 155–157, 162–165

EM Expectation-Maximization. 20, 21

EMD Earth Mover’s Distance. 21

EXIF exchangeable image file format. 16, 59, 149, 150, 155, 157, 159

FCS Flickr Context Similarity. 54, 118, 119, 121, 125–127, 129, 173, 174

FN false negative. 31–33, 46, 49, 143, 151

FP false positive. 31–33, 46, 49, 143, 151

FTS Flickr Tag Similarity. 54, 119, 121, 125–127, 129, 130, 134, 135, 143, 173, 174

GMM Gaussian Mixture Model. 18, 159

HIT Human Intelligence Task. 41, 81–85, 89, 143, 151

- HMM** Hidden Markov Model. 18
- hP** hierarchical Precision. 106
- hR** hierarchial Recall. 106
- HS** Hierarchical Score. 105, 107–109, 111, 113–116, 119, 121, 125
- HSO** Hirst and St-Onge’s measure. 52, 118, 119, 125, 127, 129, 173
- HSV** hue, saturation, and value. 17, 155
- iAP** interpolated Average Precision. 32, 47, 108, 109, 111–113, 115, 116, 152, 160–162, 165, 166
- IC** Information Content. 19, 53, 54
- ICPR** International Conference on Pattern Recognition. 147, 149, 150, 152, 153, 156, 157, 162–166
- IDMT** Institute for Digital Media Technology. 80, 150
- infAP** inferred Average Precision. 32, 46, 47, 166
- IR** Information Retrieval. iii, 2–6, 22–24, 26, 29, 30, 34–37, 39, 41, 42, 48, 55, 58, 79, 100, 116, 131, 169, 176
- JCN** Jiang and Conrath’s measure. 53, 55, 118, 119, 125–127, 129, 173
- LC** label cardinality. 31, 109, 142
- LCH** Leacock and Chodorow’s measure. 52, 118, 119, 127, 129
- LCS** least common super-concept. 52
- LD** label density. 31, 108, 109, 111, 113, 115, 138, 172
- LDA** Latent Dirichlet Allocation. 18, 21
- LDA*** Linear Discriminant Analysis. 19, 155, 159
- LESK** Lesk’s measure. 53, 118, 119, 121, 126, 127, 129
- LIN** Lin’s measure. 53, 118, 119, 121, 127, 129
- LSCOM** Large-Scale Concept Ontology for Multimedia. 27, 40, 74, 78
- MAP** Mean Average Precision. 2, 22, 32, 46, 47, 58, 132, 134, 171
- MAP-Ex** example-based Mean Average Precision. 133–135
- MiAP** Mean interpolated Average Precision. 32, 158–160, 165, 166
- MIR** Multimedia Information Retrieval. 36, 148
- MIREX** Music Information Retrieval Evaluation eXchange. 36
- mm-LDA** multi-modal Latent Dirichlet Allocation. 21

- MPEG** Moving Picture Experts Group. 16
- MRF** Markov Random Field. 18, 157
- MTurk** Amazon Mechanical Turk. 41–43, 80–83, 85, 90, 92, 128, 130, 143, 145, 146, 150, 151, 162, 166, 175, 176
- NGD** Normalised Google Distance. 54
- NIST** National Institute of Standards and Technology. 35
- NN** nearest-neighbour approach. 18, 21, 155, 157, 159
- NWD** Normalised Web Distance. 54
- OS** Ontology Score. 85, 86, 88, 89, 91, 92, 94, 99, 102–109, 111, 113–120, 123, 125, 129–131, 134, 147, 151–153, 155, 157, 158, 163, 172–174, 176
- OS-FCS** Ontology Score with Flickr Context Similarity. 147, 152, 153, 158–160, 166
- OS-FTS** Ontology Score with Flickr Tag Similarity. 131, 134, 146, 172, 176
- OWL** Web Ontology Language. 20, 70
- PCA** Principal Component Analysis. 19
- PLSA** probabilistic latent semantic analysis. 22, 93
- PR-curve** Precision-Recall Curve. 32, 46, 47
- PTO** Photo Tagging Ontology. VI, 65, 69, 70, 72–74, 78, 104–106, 116, 118, 120, 126, 129, 148, 150, 166, 171, 174
- RES** Resnik’s measure. 53, 118, 119, 121, 127, 129
- ROC** Receiver Operating Characteristics. 32, 46, 47
- ROMIP** Russian Information Retrieval Evaluation Seminar. 36
- SIFT** Scale-Invariant Feature Transform. 16, 17, 22, 155, 157, 159, 162, 166
- SR-Precision** Semantic R-Precision. 131, 134–138, 140–143, 145, 146, 172, 173, 176
- SUN** Scene UNderstanding. 29
- SVM** Support Vector Machine. 18, 20–22, 93, 94, 155, 157, 159, 166
- tf-idf** term frequency - inverse document frequency. 22
- TN** true negative. 31, 32, 46
- TP** true positive. 31, 32, 46, 143
- tr-mmLDA** topic-regression multi-modal Latent Dirichlet Allocation. 21
- TREC** Text REtrieval Conference. 27, 32, 35, 36, 42, 58, 148

TRECvid TREC Video retrieval evaluation. 1, 27, 29, 36, 148, 171

VCDT Visual Concept Detection and Annotation Task. 66, 68, 69, 78, 80, 107, 147–149, 152, 155, 165, 166, 171, 173

VEC Patwardhan's measure. 53, 118, 119, 125, 127, 129, 173

VIR Visual Information Retrieval. iii, 5, 11, 13, 23–26, 29–31, 34, 35, 37, 39, 42, 44, 55, 58, 68, 131, 169

VOC Visual Object Classes. 1, 29, 36, 39, 40, 74, 78

WIKI Milne and Witten's measure. 53, 118, 119, 121, 129

WUP Wu and Palmer's measure. 52, 118, 119, 123, 127, 129

WWW World Wide Web. 54, 118, 129, 170

xinfAP extended inferred Average Precision. 32, 47

Bibliography

- Abdel-Hakim, A. E. and A. A. Farag (2006). CSIFT: A SIFT descriptor with color invariant characteristics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Volume 2, pp. 1978–1983.
- Ah-Pine, J., S. Clinchant, G. Csurka, and Y. Liu (2009). XRCE's Participation in ImageCLEF 2009. In *Working notes for the CLEF 2009 Workshop*, Corfu, Greece.
- Alonso, O. and S. Mizzaro (2009). Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pp. 15–16.
- Alonso, O., D. E. Rose, and B. Stewart (2008). Crowdsourcing for relevance evaluation. *ACM SIGIR Forum* 42(2), 9–15.
- Ames, M. and M. Naaman (2007). Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 971–980.
- Armitage, L. and P. Enser (1997). Analysis of user need in image archives. *Journal of information science* 23(4), 287.
- Aslam, J. A., V. Pavlu, and R. Savell (2003). A unified model for metasearch, pooling, and system evaluation. In *Proceedings of the 12th international conference on Information and knowledge management (CIKM)*, pp. 491.
- Aslam, J. A., E. Yilmaz, and V. Pavlu (2005). The maximum entropy method for analyzing retrieval measures. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 27–34.
- Bailey, P., N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz (2008). Relevance Assessment : Are Judges Exchangeable and Does it Matter? In *Proceedings of the 31st ACM SIGIR conference on Research and development in information retrieval*.
- Baillie, M., L. Azzopardi, and I. Ruthven (2008, March). Evaluating epistemic uncertainty under incomplete assessments. *Information Processing & Management* 44(2), 811–837.
- Banerjee, S. and T. Pedersen (2003). Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 805–810.
- Barnard, K., P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan (2003). Matching words and pictures. *The Journal of Machine Learning Research* 3, 1107–1135.
- Batal, R. A. and P. Mulhem (2010). MRIM-LIG at ImageCLEF 2010 Visual Concept Detection and Annotation task. In *Working Notes of CLEF 2010, Padova, Italy*.

- Berg, T. L. and D. Forsyth (2006). Animals on the Web. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, Volume 2, pp. 1463–1470.
- Binder, A. and M. Kawanabe (2010). Enhancing Recognition of Visual Concepts with Primitive Color Histograms via Non-sparse Multiple Kernel Learning. In C. Peters, T. Tsirikka, H. Müller, K.-C. Jayashree, G. J. Jones, J. Gonzalo, and B. Caputo (Eds.), *Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross-Language Evaluation Forum 2009, Revised Selected Papers* (LNCS ed.), Corfu, Greece.
- Bischoff, K., C. Firan, W. Nejdl, and R. Paiu (2010). Bridging the gap between tagging and querying vocabularies: Analyses and applications for enhancing multimedia ir. *Web Semantics: Science, Services and Agents on the World Wide Web 8*(2-3), 97–109.
- Blair, D. C. (2002). Some thoughts on the reported results of TREC. *Information Processing & Management 38*(3), 445–451.
- Blei, D. M. and M. I. Jordan (2003). Modeling annotated data. In *Proceedings of the 26th ACM SIGIR conference on Research and development in informaion retrieval*, pp. 127–134.
- Blei, D. M. and J. D. McAuliffe (2007). Supervised topic models. In *Proceedings of Conference on Neural Information Processing Systems (NIPS)*.
- Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research 3*, 993–1022.
- Blockeel, H., M. Bruynooghe, S. Džeroski, J. Ramon, and J. Struyf (2002). Hierarchical Multi-Classification. In *Proceedings of the SIGKDD Workshop on Multi-Relational Data Mining*, pp. 21–35.
- Bodoff, D. (2008, May). Test theory for evaluating reliability of IR test collections. *Information Processing & Management 44*(3), 1117–1145.
- Bollegala, D., Y. Matsuo, and M. Ishizuka (2007). Measuring semantic similarity between words using web search engines. In *Proceedings of WWW*, Volume 7, pp. 757–786.
- Bosch, A., X. Munoz, and R. Marti (2007). A review : Which is the best way to organize / classify images by content? *Image and vision computing 25*(6), 778–791.
- Bosch, A., A. Zisserman, and X. Muoz (2008). Scene Classification Using a Hybrid Generative / Discriminative Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*(4), 712–727.
- Boutell, M. R., J. Luo, X. Shen, and C. M. Brown (2004). Learning multi-label scene classification. *Pattern Recognition 37*(9), 1757 – 1771.
- Brants, T. (2000). Inter-annotator agreement for a German newspaper corpus. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.
- Breitman, K. K., M. A. Casanova, and W. Truszkowski (2007). *Semantic Web: Concepts, Technologies and Applications*. Springer Verlag.
- Brennan, R. L. and D. J. Prediger (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement 41*(3), 687.

- Buckley, C. and E. M. Voorhees (2000). Evaluating evaluation measure stability. In *Proceedings of the 23rd ACM SIGIR conference on Research and development in information retrieval*, pp. 33–40.
- Buckley, C. and E. M. Voorhees (2004). Retrieval Evaluation with Incomplete Information. In *Proceedings of the 27th ACM SIGIR conference on Research and development in information retrieval*, pp. 25–32.
- Budanitsky, A. and G. Hirst (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics* 32(1), 13–47.
- Cai, L. (2008). *Multilabel classification over category taxonomies*. Ph. D. thesis, Brown University.
- Cai, L. and T. Hofmann (2007). Exploiting Known Taxonomies in Learning Overlapping Concepts. In *Proceedings of International Joint Conferences on Artificial Intelligence*, pp. 714–719.
- Carneiro, G., A. B. Chan, P. J. Moreno, and N. Vasconcelos (2007). Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(3), 394–410.
- Carneiro, G. and N. Vasconcelos (2005). Formulating semantic image annotation as a supervised learning problem. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Volume 2, pp. 163–168.
- Carterette, B. (2009). On rank correlation and the distance between rankings. In *Proceedings of the 32nd ACM SIGIR conference on Research and development in information retrieval*, pp. 436–443.
- Carterette, B. A. (2008, November). *Low-cost and robust evaluation of information retrieval systems*. Phd thesis, University of Massachusetts Amherst.
- Cesa-Bianchi, N., C. Gentile, and L. Zaniboni (2006). Incremental algorithms for hierarchical classification. *The Journal of Machine Learning Research* 7, 31–54.
- Chen, J. J., N. J. Menezes, and A. D. Bradley (2011). Opportunities for Crowdsourcing Research on Amazon Mechanical Turk. *Interfaces* 5, 3.
- Chen, K.-T., C.-C. Wu, Y.-C. Chang, and C.-L. Lei (2009). A crowdsourcable QoE evaluation framework for multimedia content. In *Proceedings of the 17th ACM International Conference on Multimedia*.
- Chen, X., Y. Mu, S. Yan, and T. S. Chua (2010). Efficient large-scale image annotation by probabilistic collaborative multi-label propagation. In *Proceedings of the ACM international conference on Multimedia*, pp. 35–44.
- Chklovski, T. and R. Mihalcea (2003). Exploiting agreement and disagreement of human annotators for word sense disambiguation. In *Proceedings of RANLP*.
- Chua, T.-s., J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng (2009). NUS-WIDE : A Real-World Web Image Database from National University of Singapore. In *International Conference on Image and Video Retrieval (CIVR)*.
- Cilibrasi, R. and P. Vitanyi (2007). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering* 19(3), 370–383.

- Cleverdon, C., J. Mills, and M. Keen (1966). Factors determining the performance of indexing systems. *Aslib Cranfield Research Project*.
- Cleverdon, C. W. (1967). The Cranfield tests on index language devices. *Aslib proceedings* 19(6), 173–193.
- Clough, P., H. Müller, and M. Sanderson (2005). The CLEF 2004 Cross-Language Image Retrieval Track. In C. Peters, P. D. Clough, G. J. F. Jones, J. Gonzalo, M. Kluck, and B. Magnini (Eds.), *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Revised Selected Papers*, Volume 3491, pp. 597–613. LNCS.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1), 37–46.
- Cordi, V., P. Lombardi, M. Martelli, and V. Mascardi (2005). An Ontology-Based Similarity between Sets of Concepts. In *Proceedings of WOA, Italy*, pp. 16–21.
- Cormack, G. V. and T. R. Lynam (2006). Statistical precision of information retrieval evaluation. In *Proceedings of the 29th ACM SIGIR conference on Research and development in information retrieval*, pp. 533–540.
- Cormack, G. V. and T. R. Lynam (2007). Validity and power of t-test for comparing MAP and GMAP. In *Proceedings of the 30th ACM SIGIR conference on Research and development in information retrieval*, pp. 753–754.
- Cormack, G. V., C. R. Palmer, and C. L. A. Clarke (1998). Efficient construction of large test collections. In *Proceedings of the 21st ACM SIGIR conference on Research and development in information retrieval*, pp. 282–289.
- Costa, E., A. Lorena, A. Carvalho, and A. Freitas (2007). A Review of Performance Evaluation Measures for Hierarchical Classifiers. In *Proceedings of the AAAI 2007 workshop: Evaluation methods for machine learning*, pp. 1–6.
- Cuadra, C. A. and R. V. Katter (1967). Opening the black box of relevance. *Journal of Documentation* 23(4), 291–303.
- D'Amato, C. (2007). *Similarity-based Learning Methods for the Semantic Web*. Phd, University of Bari.
- Daróczy, B., I. Petrás, A. A. Benczúr, Z. Fekete, D. Nemeskey, D. Siklósi, and Z. Weiner (2010). Interest Point and Segmentation-Based Photo Annotation. In C. Peters, T. Tsirikia, H. Müller, J. Kalpathy-Cramer, J. F. G. Jones, J. Gonzalo, and B. Caputo (Eds.), *Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross Language Evaluation Forum 2009, Revised Selected Papers*, Corfu, Greece. LNCS.
- Daróczy, B., I. Petrás, A. A. Benczúr, D. Nemeskey, and R. Pethes (2010). SZTAKI @ ImageCLEF 2010. In *Working Notes of CLEF 2010, Padova, Italy*.
- Datta, R., D. Joshi, J. Li, and J. Wang (2006). Studying aesthetics in photographic images using a computational approach. In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 288–301.

- Datta, R., D. Joshi, J. Li, and J. Z. Wang (2007). Tagging over time: Real-world image annotation by lightweight meta-learning. In *Proceedings of the 15th international Conference on Multimedia*, pp. 393–402.
- Datta, R., D. Joshi, J. Li, and J. Z. Wang (2008). Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Transactions on Computing Surveys* 40(2), 1–60.
- de Vries, A. P., G. Kazai, and M. Lalmas (2004). Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In *RIAO Conference Proceedings*, pp. 463–473.
- Demartini, G. and S. Mizzaro (2006). A classification of IR effectiveness metrics. In *Advances in Information Retrieval*, Volume 3936, pp. 488–491. LNCS.
- Deng, J., A. C. Berg, K. Li, and L. Fei-Fei (2010). What does classifying more than 10,000 image categories tell us? *European Conference on Computer Vision (ECCV) 6315*, 71–48.
- Deng, J., W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei (2009). ImageNet: a large-scale hierarchical image database. In *Proceedings of CVPR*, pp. 710–719.
- Depeursinge, A. and H. Müller (2010). Fusion Techniques for Combining Textual and Visual Information Retrieval. In W. B. Croft, H. Müller, P. Clough, T. Deselaers, and B. Caputo (Eds.), *ImageCLEF Experimental Evaluation in Visual Information Retrieval*, Volume 32 of *The Information Retrieval Series*, Chapter 6, pp. 95–114. Springer.
- Dimitrovski, I., D. Kocev, S. Loskovska, and S. Džeroski (2010). Detection of Visual Concepts and Annotation of Images using Predictive Clustering Trees. In *Working Notes of CLEF 2010, Padova, Italy*.
- Donmez, P., J. G. Carbonell, and J. Schneider (2009). Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM conference on Knowledge discovery and data mining (SIGKDD)*, pp. 259–268.
- Douze, M., M. Guillaumin, T. Mensink, C. Schmid, and J. Verbeek (2009). INRIA-LEARs participation to ImageCLEF 2009. In *Working Notes for the CLEF 2009 Workshop, Corfu, Greece*.
- Downie, J. S. (2008). The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology* 29(4), 247–255.
- Duan, M., A. Ulges, T. M. Breuel, and X. Wu (2009). Style Modeling for Tagging Personal Photo Collections. In *Proceeding of the ACM Conference on Image and Video Retrieval (CIVR)*.
- Dumont, E., Z.-Q. Zhao, H. Glotin, and S. Paris (2010). A new TFIDF Bag of Visual Words for Concept Detection. In C. Peters, T. Tsirikia, H. Müller, J. Kalpathy-Cramer, J. F. G. Jones, J. Gonzalo, and B. Caputo (Eds.), *Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross Language Evaluation Forum 2009, Revised Selected Papers*, Corfu, Greece. LNCS.
- Dupret, G. and B. Piwowarski (2010). A User Behavior Model for Average Precision and its Generalization to Graded Judgments. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 531–538.

- Duygulu, P., K. Barnard, J. F. G. Freitas, and D. A. Forsyth (2002). Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In *Proceedings of the 7th European Conference on Computer Vision (ECCV)*, pp. 97–112.
- Eakins, J. and M. Graham (1999). Content-based image retrieval. *JISC Technology Applications Programme Report 39*, 1–59.
- Efron, M. (2009). Using multiple query aspects to build test collections without human relevance judgments. In *Advances in Information Retrieval*, Volume 5478, pp. 276–287. LNCS.
- Eickhoff, C. and A. P. de Vries (2011). How Crowdsourcable is Your Task? In *Workshop on Crowdsourcing for Search and Data Mining*.
- Endres, I., A. Farhadi, D. Hoiem, and D. A. Forsyth (2010). The benefits and challenges of collecting richer object annotations. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1–8.
- Enser, P. (2000). Visual image retrieval: seeking the alliance of concept-based and content-based paradigms. *Journal of Information Science* 26(4), 199.
- Escalante, H. J., J. A. Gonzalez, C. A. Hernandez, A. Lopez, M. Montex, E. Morales, E. Ruiz, L. E. Sucar, and L. Villaseñor (2009). TIA-INAOE's Participation at ImageCLEF 2009. In *Working Notes for the CLEF 2009 Workshop, Corfu, Greece*.
- Escalante, H. J., C. A. Hernández, J. A. Gonzalez, A. López-López, M. Montes, E. F. Morales, L. Enrique Sucar, L. Villaseñor, and M. Grubinger (2010). The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding* 114(4), 419–428.
- Everingham, M., L. van Gool, C. K. I. Williams, J. Winn, and A. Zisserman (2010). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88(2), 303–538.
- Everingham, M., A. Zisserman, C. K. I. Williams, L. van Gool, M. Allan, and E. al. (2006). The 2005 PASCAL Visual Object Classes Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment (PASCAL Workshop 05)*, Number 3944 in LNAI, Southampton, UK, pp. 117–176.
- Fakeri-Tabrizi, A., S. Tollari, N. Usunier, M.-R. Amini, and P. Gallinari (2010). UPMC/LIP6 at ImageCLEFannotation 2010. In *Working Notes of CLEF 2010, Padova, Italy*.
- Fakeri-Tabrizi, A., S. Tollari, N. Usunier, and P. Gallinari (2010). Improving Image Annotation in Imbalanced Classification Problems with Ranking SVM. In C. Peters, T. Tsikrika, H. Müller, J. Kalpathy-Cramer, J. F. G. Jones, J. Gonzalo, and B. Caputo (Eds.), *Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross Language Evaluation Forum 2009, Revised Selected Papers*, Corfu, Greece. LNCS.
- Fan, J., Y. Gao, and H. Luo (2004). Multi-level annotation of natural scenes using dominant image components and semantic concepts. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 540–547.
- Fan, J., Y. Gao, and H. Luo (2007). Hierarchical classification for automatic image annotation. In *Proceedings of the 30th international ACM SIGIR conference on Research and development in information retrieval*, pp. 111–118.

- Fan, J., Y. Gao, H. Luo, and S. Satoh (2008). New Approach for Hierarchical Classifier Training and Multi-level Image Annotation. In *Advances in Multimedia Modeling*, Volume 4903, pp. 45–57. LNCS.
- Fellbaum, C. (1998). *WordNet: An electronic lexical database*. The MIT press.
- Felstiner, A. L. (2010). Working the crowd: employment and labor law in the crowdsourcing industry. Technical report, Available at <http://ssrn.com/abstract=1593853>.
- Feng, D., S. Besana, and R. Zajac (2009). Acquiring High Quality Non-Expert Knowledge from On-demand Workforce. In *Proceedings of the Workshop: The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pp. 51–56.
- Feng, S. L., R. Manmatha, and V. Lavrenko (2004). Multiple bernoulli relevance models for image and video annotation. In *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society.
- Ferecatu, M. and H. Sahbi (2009). TELECOM ParisTech at ImageClef 2009: Large Scale Visual Concept Detection and Annotation Task. In *Working Notes of the CLEF 2009 Workshop, Corfu, Greece*.
- Fergus, R., L. Fei-Fei, P. Perona, and A. Zisserman (2005). Learning Object Categories from Google's Image Search. In *Tenth IEEE International Conference on Computer Vision (ICCV)*, pp. 1816–1823.
- Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppim (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pp. 406–414.
- Fort, K., G. Adda, and K. B. Cohen (2011). Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics* 37(2), 413–420.
- Freitas, A. A. and A. C. de Carvalho (2007). A tutorial on hierarchical classification with applications in bioinformatics. *Research and Trends in Data Mining Technologies and Applications*, 175–208.
- Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Academic Press.
- Gabrilovich, E. and S. Markovitch (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Volume 7, pp. 1606–1611. Morgan Kaufmann Publishers Inc.
- Gibbons, J. and S. Chakraborti (2003). *Nonparametric statistical inference*, Volume 168. CRC Press.
- Glotin, H., A. Fakeri-Tabrizi, P. Mulhem, M. Ferecatu, Z. Zhao, S. Tollari, G. Quenot, H. Sahbi, E. Dumont, and P. Gallinari (2009). Comparison of Various AVEIR Visual Concept Detectors with an Index of Carefulness. In *Working Notes of the CLEF 2009 Workshop, Corfu, Greece*.
- Gomez-Perez, A., M. Fernández-López, and O. Corcho (2003). *Ontological engineering*. Springer Verlag.
- Gracia, J. and E. Mena (2008). Web-based Measure of Semantic Relatedness. In *Proceedings of the International Conference on Web Information Systems Engineering*.

- Grady, C. and M. Lease (2010). Crowdsourcing document relevance assessment with Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 172–179.
- Greisdorf, H. (2000). Relevance: An Interdisciplinary and Information Science Perspective. *Informing Science* 3, 67–71.
- Greisdorf, H. and B. O Connor (2002). Modelling what users see when they look at images: a cognitive viewpoint. *Journal of Documentation* 58(1), 6–29.
- Griffin, G., A. Holub, and P. Perona (2007). Caltech-256 Object Category Dataset. Technical report, California Institute of Technology.
- Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies* 43(5), 907–928.
- Gruber, T. R. (2009). Ontology. In L. Liu and O. M. Tamer (Eds.), *Encyclopedia of Database Systems*. Springer Verlag.
- Grubinger, M. (2007). *Analysis and Evaluation of Visual Information Systems Performance*. Ph. D. thesis, Victoria University.
- Grubinger, M., P. Clough, H. Müller, and T. Deselaers (2006). The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems. In *Proceedings of the International Workshop OntoImage'2006*, pp. 13–23.
- Guarino, N. (1998). Formal ontology in information systems. In *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems (FOIS)*, Trento, Italy, pp. 3–15.
- Guillaumin, M., T. Mensink, J. Verbeek, and C. Schmid (2009). TagProp : Discriminative Metric Learning in Nearest Neighbor Models for Image Auto-Annotation. In *IEEE 12th International Conference on Computer Vision (ICCV)*, pp. 309–316.
- Gupta, M., R. Li, Z. Yin, and J. Han (2010). Survey on social tagging techniques. *SIGKDD Explorations Newsletter* 12(1), 58–72.
- Hanbury, A. (2007). A study of vocabularies for image annotation. In *Semantic Multimedia*, Volume 4816, pp. 284–287. LNCS.
- Hare, J. S. and P. H. Lewis (2009). IAM@ImageCLEFPhotoAnnotation 2009: Naive application of a linear-algebraic semantic space. In *Working Notes of the CLEF 2009 Workshop, Corfu, Greece*.
- Harman, D. (2011, May). Information Retrieval Evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 3(2), 1–119.
- Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science* 47(1), 37–49.
- Haubold, A. and A. Natsev (2008). Web-based information content and its application to concept-based video retrieval. In *Proceedings of the 2008 international conference on Content-based image and video retrieval (CIVR)*, pp. 437–446.
- Hauptmann, A. G., M. G. Christel, and R. Yan (2008). Video retrieval based on semantic concepts. *Proceedings of the IEEE* 96(4), 602–622.

- Hirst, G. and D. St-Onge (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: A Lexical Database for English* 305, 332.
- Hitzler, P., M. Krötzsch, and S. Rudolph (2009). *Foundations of semantic web technologies*. Chapman & Hall/CRC.
- Hollink, L., A. T. Schreiber, B. J. Wielinga, and M. Worrying (2004). Classification of user image descriptions. *International Journal of Human-Computer Studies* 61(5), 601–626.
- Horton, J. J. and R. J. Zeckhauser (2010). Algorithmic Wage Negotiations: Applications to Paid Crowdsourcing. In *Proceedings of CrowdConf*. CrowdConf.
- Howe, J. (2006). Crowdsourcing - A Definition. <http://crowdsourcing.typepad.com/cs/2006/06/>, last accessed 24.11.2009.
- Hsueh, P. Y., P. Melville, and V. Sindhwani (2009). Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*.
- Huang, X. and Y. Zhou (2010). An Asymmetric Similarity Measure for Tag Clustering on Flickr. In *Proceedings of International Asia-Pacific Web Conference*, pp. 171–177.
- Huiskes, M. J. and M. S. Lew (2008). The MIR Flickr Retrieval Evaluation. In *Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval (MIR)*.
- Huiskes, M. J., B. Thomee, and M. S. Lew (2010). New Trends and Ideas in Visual Concept Detection The MIR Flickr Retrieval Evaluation Initiative. In *International Conference on Multimedia Information Retrieval*, pp. 527–536.
- Hwang, S. J. and K. Grauman (2010). Reading between the lines: Object localization using implicit cues from image tags. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2971–2978.
- Iftene, A., L. Vamanu, and C. Croitoru (2010). UAIC at ImageCLEF 2009 Photo Annotation Task. In C. Peters, T. Tsirikika, H. Müller, J. Kalpathy-Cramer, J. F. G. Jones, J. Gonzalo, and B. Caputo (Eds.), *Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross-Language Evaluation Forum 2009, Revised Selected Papers*, Corfu, Greece. LNCS.
- Ipeirotis, P. G. (2010). Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students* 17(2), 16–21.
- Ipeirotis, P. G., L. Gravano, and M. Sahami (2001). Probe, count, and classify: categorizing hidden web databases. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, pp. 67–78.
- Ipeirotis, P. G., F. Provost, and J. Wang (2010). Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pp. 64–67.
- ISO2788 (1986). Guidelines for the establishment and development of monolingual thesauri. Technical report, International Organization for Standardization (ISO).
- Jain, A. K., R. P. W. Duin, and J. Mao (2000). Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence* 22(1), 4–37.

- Järvelin, K. and J. Kekäläinen (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20(4), 446.
- Jeon, J., V. Lavrenko, and R. Manmatha (2003). Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th ACM SIGIR conference on Research and development in informaion retrieval*, pp. 119–126.
- Jiang, J. J. and D. W. Conrath (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of Conference on Research on Computational Linguistics*.
- Jiang, Y. G., C. W. Ngo, and S. F. Chang (2009). Semantic context transfer across heterogeneous sources for domain adaptive video search. In *Proceedings of the ACM International Conference on Multimedia*, pp. 155–164.
- Jiang, Y. G., J. Yang, C. W. Ngo, and A. G. Hauptmann (2010). Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia* 12(1), 42–53.
- Jin, Y., L. Khan, L. Wang, and M. Awad (2005). Image annotations by combining multiple evidence & wordNet. In *Proceedings of the 13th ACM international conference on Multimedia*, pp. 706–715.
- Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer.
- Jørgensen, C. (1998). Attributes of images in describing tasks. *Information Processing & Management* 34(2-3), 161–174.
- Kazai, G. and N. Milic-Frayling (2009). On the Evaluation of the Quality of Relevance Assessments Collected through Crowdsourcing. In *SIGIR 2009 Workshop on the Future of IR Evaluation*.
- Ke, Y., X. Tang, and F. Jing (2006). The design of high-level features for photo quality assessment. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 419–426.
- Kendall, M. (1938). A New Measure of Rank Correlation. *Biometrika* 30, 81–89.
- Kennedy, L., A. Hauptmann, M. Naphade, J. R. Smith, and S.-F. Chang (2006). LSCOM Lexicon Definitions and Annotations Version 1.0. In *DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report*.
- King, M. (2003). Living up to standards. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: Are evaluation methods, metrics and resources reusable?*, pp. 65–72.
- Kiritchenko, S. (2005). *Hierarchical text categorization and its application to bioinformatics*. Ph. D. thesis, University of Ottawa.
- Kiritchenko, S., S. Matwin, and F. Famili (2005). Functional Annotation of Genes Using Hierarchical Text Categorization. In *Proceedings of the ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*.
- Kirk, D., A. Sellen, C. Rother, and K. Wood (2006). Understanding photowork. In *Proceedings of the ACM SIGCHI conference on Human Factors in computing systems*, pp. 761–770.
- Kolmogoroff, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari* 4(1), 83–91.

- Kuhn, H. (1955). The Hungarian method for the assignment problem. *Naval research logistics quarterly* 2(1-2), 83–97.
- Kumar, R. and S. Vassilvitskii (2010). Generalized distances between rankings. In *Proceedings of the 19th international conference on World wide web*, pp. 571–580.
- Kuylen, A. and T. Verhallen (1988). The natural grouping of banks: a new methodology for positioning research. *Practical contributions of research*, 219–234.
- Laine-Hernandez, M. and S. Westman (2006). Image semantics in the description and categorization of journalistic photographs. *Proceedings of the American Society for Information Science and Technology* 43(1), 1–25.
- Landis, J. R. and G. G. Koch (1977). The measurement of observer agreement for categorical data. *Biometrics* 33(1), 159–174.
- Lavrenko, V., R. Manmatha, and J. Jeon (2003). A Model for Learning the Semantics of Pictures. In *Proceedings of Conference on Neural Information Processing Systems (NIPS)*.
- Leacock, C. and M. Chodorow (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. *WordNet: A Lexical Reference System and its Application* 49(2), 265–283.
- Lenat, D. B. and R. V. Guha (1989). *Building large knowledge-based systems: representation and inference in the Cyc project*. Addison Wesley Publishing Company.
- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the international Conference on Systems Documentation*, pp. 24–26.
- Lesk, M. E. and G. Salton (1968). Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval* 4(4), 343–359.
- Lew, M. S., N. Sebe, C. Djeraba, and R. Jain (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)* 2(1), 1–19.
- Li, J. and J. Z. Wang (2007). Real-time computerized annotation of pictures. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(6), 985–1002.
- Li, W., J. Min, and G. J. F. Jones (2010). A Text-Based Approach to the ImageCLEF 2010 Photo Annotation Task. In *Working Notes of CLEF 2010, Padova, Italy*.
- Li, X., C. Wu, C. Zach, S. Lazebnik, and J. M. Frahm (2008). Modeling and Recognition of Landmark Image Collections Using Iconic Scene Graphs. In *Proceedings of the 10th European Conference on Computer Vision (ECCV)*, pp. 427–440.
- Lin, C. Y., B. L. Tseng, and J. R. Smith (2003). Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. In *Proceedings of the TRECVID 2003 Workshop*.
- Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the international Conference on Machine Learning*.

- Lindsey, R., V. D. Veksler, A. Grintsveyg, and W. D. Gray (2007). Be wary of what your computer reads: the effects of corpus selection on measuring semantic relatedness. In *8th International Conference of Cognitive Modeling (ICCM)*.
- Ling, X., J. Jia, N. Yu, and M. Li (2008). Tagrank-Measuring tag importance for image annotation. In *International Conference on Multimedia and Expo (ICME)*, pp. 109–112.
- Little, S., A. Llorente, and S. Rüger (2010). An Overview of Evaluation Campaigns in Multimedia Retrieval. In H. Müller, P. Clough, T. Deselaers, and B. Caputo (Eds.), *ImageCLEF - Experimental Evaluation of Visual Information Retrieval*, The Information Retrieval Series, Chapter 27, pp. 507–525. Springer.
- Liu, D., X.-s. Hua, M. Wang, and H. Zhang (2009). Boost search relevance for tag-based social image retrieval. In *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1636–1639.
- Liu, D., S. Yan, Y. Rui, and H. J. Zhang (2010). Unified tag analysis with multi-edge graph. In *Proceedings of the international conference on Multimedia*, pp. 25–34.
- Liu, H. and P. Singh (2004). ConceptNet-a practical commonsense reasoning tool-kit. *BT technology journal* 22(4), 211–226.
- Liu, J., B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma (2007). Dual cross-media relevance model for image annotation. In *Proceedings of the 15th international conference on Multimedia*, pp. 605–614.
- Liu, Y. and E. Shriberg (2007). Comparing Evaluation Metrics for Sentence Boundary Detection. In *IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Volume 4, pp. 185188.
- Liu, Y., D. Zhang, G. Lu, and W. Y. Ma (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition* 40(1), 262–282.
- Llorente, A., E. Motta, and S. Rüger (2009). Image annotation refinement using Web-based keyword correlation. In *Semantic Multimedia*, Volume 5887, pp. 188–191. LNCS.
- Llorente, A., E. Motta, and S. Rüger (2010). Exploring the Semantics Behind a Collection to Improve Automated Image Annotation. In C. Peters, T. Tsirikia, H. Müller, J. Kalpathy-Cramer, J. F. G. Jones, J. Gonzalo, and B. Caputo (Eds.), *Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross Language Evaluation Forum 2009, Revised Selected Papers*, Corfu, Greece. LNCS.
- Lord, P. W., R. D. Stevens, A. Brass, and C. A. Goble (2003). Investigating semantic similarity measures across the Gene Ontology: The relationship between sequence and annotation. *Bioinformatics* 19(10), 1275–1283.
- Loui, A., J. Luo, S. F. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, and A. Yanagawa (2007). Kodak’s consumer video benchmark data set: concept definition and annotation. In *Proceedings of the international workshop on multimedia information retrieval*, pp. 245–254.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60(2), 91–110.
- Makadia, A., V. Pavlovic, and S. Kumar (2008). A new baseline for image annotation. In *European Conference on Computer Vision (ECCV)*, Volume 8, pp. 316–329.

- Makadia, A., V. Pavlovic, and S. Kumar (2010). Baselines for Image Annotation. *International Journal of Computer Vision* 90(1), 88–105.
- Manjunath, B. S., P. Salembier, and T. Sikora (2002). *Introduction to MPEG-7: multimedia content description interface*. John Wiley & Sons Inc.
- Manning, C. D., P. Raghavan, and H. Schütze (2009, April). *An Introduction to Information Retrieval [Draft]*. Cambridge University Press, UK.
- Markkula, M. and E. Sormunen (1998). Searching for photos—journalists’ practices in pictorial ir. In *The Challenge of Image Retrieval, A Workshop and Symposium on Image Retrieval, Electronic Workshops in Computing, Newcastle upon Tyne*, Volume 56.
- Martin, D., C. Fowlkes, D. Tal, and J. Malik (2001). A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *Proceedings 8th international Conference on Computer Vision*, pp. 416–423.
- Mason, W. and D. J. Watts (2009). Financial Incentives and the "Performance of Crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pp. 77–85.
- Maynard, D., W. Peters, and Y. Li (2006). Metrics for evaluation of ontology-based information extraction. In *WWW 2006 Workshop on Evaluation of Ontologies for the Web (EON), Edinburgh, Scotland*.
- Mbanya, E., C. Hentschel, S. Gerke, M. Liu, A. Nürnberger, and P. Ndjiki-Nya (2010). Augmenting Bag-of-Words - Category Specific Features and Concept Reasoning. In *Working Notes of CLEF 2010, Padova, Italy*.
- Mei, T., Y. Wang, X. S. Hua, S. Gong, and S. Li (2008). Coherent image annotation by learning semantic distance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Melucci, M. (2007). On rank correlation in information retrieval evaluation. *ACM SIGIR Forum* 41(1), 18–33.
- Mensink, T., G. Csurka, F. Perronnin, J. Sánchez, and J. Verbeek (2010). LEAR and XRCE’s participation to Visual Concept Detection Task - ImageCLEF 2010. In *Working Notes of CLEF 2010, Padova, Italy*.
- Miller, A. D. and W. K. Edwards (2007). Give and take: a study of consumer photo-sharing culture and practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 347–356.
- Miller, G. A. and W. G. Charles (1991). Contextual correlates of semantic similarity. *Language and cognitive processes* 6(1), 1–28.
- Milne, D. and I. H. Witten (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the AAAI Workshop on Wikipedia and AI*.
- Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science* 48(9), 810–832.
- Moellic, P.-A. and C. Fluhr (2006). ImageEVAL 2006 Official Campaign. Technical report, ImageEVAL.
- Moffat, A. and J. Zobel (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27(1), 1–27.

- Monay, F. and D. Gatica-Perez (2004). PLSA-based image auto-annotation: constraining the latent space. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 348–351.
- Mori, Y., H. Takahashi, and R. Oka (1999). Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of 1st International Workshop on Multimedia Intelligent Storage and Retrieval Management*.
- Motohashi, N., R. Izawa, and T. Takagi (2010). Meiji University at the ImageCLEF2010 Visual Concept Detection and Annotation Task: Working notes. In *Working Notes of CLEF 2010, Padova, Italy*.
- Müller, H., P. Clough, T. Deselaers, and B. Caputo (Eds.) (2010). *ImageCLEF - Experimental Evaluation of Visual Information Retrieval*. The Information Retrieval Series. Springer.
- Müller, H., T. Deselaers, M. Grubinger, P. Clough, A. Hanbury, and W. Hersh (2007). Problems with Running a Successful Multimedia Retrieval Benchmark. In *Proceedings of the third MUSCLE/ImageCLEF workshop on image and video retrieval evaluation*.
- Müller, H., S. Marchand-Maillet, and T. Pun (2002). The Truth about Corel-Evaluation in Image Retrieval. In *Image and Video Retrieval*, Volume 2383, pp. 38–49. LNCS.
- Müller, H., W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun (2001). Performance evaluation in content-based image retrieval: overview and proposals. *Pattern Recognition Letters* 22(5), 593–601.
- Nagel, K. (2011). Multimodale Detektion und Annotation von Konzepten in Fotos. Master's thesis, Technical University of Ilmenau.
- Naphade, M., J. R. Smith, J. Tesic, S. F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis (2006). Large-scale concept ontology for multimedia. *IEEE Multimedia* 13(3), 86–91.
- Narasimhalu, A. D., M. S. Kankanhalli, and J. Wu (1997). Benchmarking multimedia databases. *Multimedia Tools and Applications* 4(3), 333–356.
- Ngiam, J. and H. Goh (2010). Learning Global and Regional Features for Photo Annotation. In C. Peters, T. Tsirikia, H. Müller, J. Kalpathy-Cramer, J. F. G. Jones, J. Gonzalo, and B. Caputo (Eds.), *Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross Language Evaluation Forum 2009, Revised Selected Papers*, Corfu, Greece. LNCS.
- Oliva, A., W. Hospital, and L. Ave (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision* 42(3), 145–175.
- Paris, S. and H. Glotin (2010). Linear SVM for LSIS Pyramidal Multi-Level Visual only Concept Detection in CLEF 2010 Challenge. In *Working Notes of CLEF 2010, Padova, Italy*.
- Patwardhan, S. (2003). Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness. Master's thesis, University of Minnesota.
- Pham, T., L. Maisonnasse, P. Mulhem, J.-P. Chevallet, G. Quénot, and R. Al Batal (2010). MRIM-LIG at ImageCLEF 2009: Robot Vision, Image annotation and retrieval tasks. In C. Peters, T. Tsirikia, H. Müller, J. Kalpathy-Cramer, J. F. G. Jones, J. Gonzalo, and B. Caputo

- (Eds.), *Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross-Language Evaluation Forum 2009, Revised Selected Papers*, Corfu, Greece. LNCS.
- Picard, R. W. and T. P. Minka (1995). Vision texture for annotation. *Multimedia systems* 3(1), 3–14.
- Putthividhy, D., H. T. Attias, and S. S. Nagarajan (2010). Topic regression multi-modal Latent Dirichlet Allocation for image annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3408–3415.
- Qi, G. J., X. S. Hua, and H. J. Zhang (2009). Learning semantic distance from community-tagged media collection. In *Proceedings of the seventeen ACM international conference on Multimedia*, pp. 243–252.
- Randolph, J. (2008). Online Kappa Calculator. <http://justus.randolph.name/kappa>, last accessed 17.12.2009.
- Randolph, J. J. (2005). Free-Marginal Multirater Kappa (multirater κ free): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. In *Joensuu Learning and Instruction Symposium*.
- Rasche, C. and C. Vertan (2010). A Novel Structural-Description Approach for Image Retrieval. In *Working Notes of CLEF 2010, Padova, Italy*.
- Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 448–453.
- Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence* 11, 95–130.
- Robertson, S. E., E. Kanoulas, and E. Yilmaz (2010). Extending average precision to graded relevance judgments. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 603–610.
- Rodden, K. and K. R. Wood (2003). How do people manage their digital photographs? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 409–416.
- Ross, J., L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, pp. 2863–2872.
- Rubenstein, H. and J. B. Goodenough (1965). Contextual correlates of synonymy. *Communications of the ACM* 8(10), 627–633.
- Rüger, S. (2009, January). Multimedia Information Retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 1(1), 1–171.
- Rui, Y., T. S. Huang, and S. F. Chang (1999). Image Retrieval: Current Techniques, Promising Directions, and Open Issues. *Journal of visual communication and image representation* 10(1), 39–62.
- Russell, B. C., A. Torralba, K. P. Murphy, and W. T. Freeman (2008). LabelMe : a database and web-based tool for image annotation. *International Journal of Computer Vision* 77(1), 157–173.

- Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology* 39(6), 1161.
- Sahami, M. and T. D. Heilman (2006). A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web*, pp. 377–386.
- Sahbi, H. and X. Li (2010). TELECOM ParisTech at ImageCLEF 2010 Photo Annotation Task: Combining Tags and Visual Features for Learning-Based Image Annotation. In *Working Notes of CLEF 2010, Padova, Italy*.
- Sakai, T. (2006). Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th ACM SIGIR conference on Research and development in information retrieval*, pp. 525–532.
- Sakai, T. (2007). On the reliability of information retrieval metrics based on graded relevance. *Information processing & management* 43(2), 531–548.
- Sakai, T. (2008). Comparing metrics across TREC and NTCIR: the robustness to system bias. In *Proceeding of the 17th ACM conference on Information and knowledge management (CIKM)*, pp. 581–590.
- Salton, G. (1971). The SMART retrieval system: experiments in automatic document processing. *Prentice-Hall, Inc. Upper Saddle River, NJ, USA*.
- Salton, G. (1992). The state of retrieval system evaluation. *Information Processing & Management* 28(4), 441–449.
- Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval* 4(4), 247–375.
- Sanderson, M. and H. Joho (2004). Forming test collections with no system pooling. In *Proceedings of the 27th ACM SIGIR conference on Research and development in information retrieval*, pp. 33–40.
- Sanderson, M. and J. Zobel (2005). Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th ACM SIGIR conference on Research and development in information retrieval*, pp. 162–169.
- Sarin, S. and W. Kameyama (2009). Joint Contribution of Global and Local Features for Image Annotation. In *Working Notes for the CLEF 2009 Workshop, Corfu, Greece*.
- Schamber, L. (1994). Relevance and Information Behavior. In *Annual review of information science and technology (ARIST)*, Volume 29, pp. 3–48.
- Schapire, R. E. and Y. Singer (2000). BoosTexter: A boosting-based system for text categorization. *Machine learning* 39(2), 135–168.
- Schroff, F., A. Criminisi, and A. Zisserman (2007). Harvesting Image Databases from the Web. In *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV)*, pp. 1–8.
- Shen, X., M. Boutell, J. Luo, and C. Brown (2004). Multi-label machine learning and its application to semantic scene classification. In *International Symposium on Electronic Imaging*.

- Sheng, V. S., F. Provost, and P. G. Ipeirotis (2008). Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proc. of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*.
- Shevade, B. and H. Sundaram (2006). A visual annotation framework using common-sensical and linguistic relationships for semantic media retrieval. In *Adaptive Multimedia Retrieval: User, Context, and Feedback: third international workshop, AMR 2005, Glasgow, UK, July 28-29, 2005: revised selected papers*, pp. 251 – 265.
- Sieweke, T. (2010). Wie organisieren und verwalten Nutzer ihre Photo-Sammlungen? Definition und Analyse von visuellen Konzepten anhand von Probandentests. Master's thesis, Technical University of Ilmenau.
- Sigurbjörnsson, B. and R. van Zwol (2008). Flickr tag recommendation based on collective knowledge. In *Proceeding of the 17th international conference on World Wide Web*, pp. 327–336.
- Silla, C. N. and A. A. Freitas (2010, April). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery* 22(1–2), 31–72.
- Smeaton, A. F., P. Over, and W. Kraaij (2006). Evaluation campaigns and TRECVID. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pp. 321–330. ACM Press.
- Smeulders, A. W. M., M. Worring, S. Santini, A. Gupta, and R. Jain (2000). Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12), 1349–1380.
- Smith, J. R. (1998). Image retrieval evaluation. In *IEEE Workshop on Content-based Access of Image and Video Libraries*, pp. 4–5.
- Snoek, C. G. M. and M. Worring (2009). Concept-based video retrieval. *Foundations and Trends in Information Retrieval* 2(4), 215–322.
- Snoek, C. G. M., M. Worring, J. C. van Gemert, J. M. Geusebroek, and A. W. M. Smeulders (2006). The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pp. 421–430.
- Snow, R., B. O'Connor, D. Jurafsky, and A. Y. Ng (2008). Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Soboroff, I., C. Nicholas, and P. Cahan (2001). Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 66–73.
- Soleymani, M. and M. Larson (2010). Crowdsourcing for Affective Annotation of Video: Development of a Viewer-reported Boredom Corpus. In V. Carvalho, M. Lease, and E. Yilmaz (Eds.), *Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, Geneva, Switzerland.
- Sorokin, A. and D. Forsyth (2008). Utility data annotation with Amazon Mechanical Turk. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*.

- Spärck Jones, K. and C. J. van Rijsbergen (1975). Report on the need for and provision of an ideal information retrieval test collection. *British Library Research and Development Report No. 5266, Computer Laboratory, University of Cambridge*, 43.
- Srihari, R. K. (1995). Automatic indexing and content-based retrieval of captioned images. *Computer* 28(9), 49–56.
- Srikanth, M., J. Varner, M. Bowden, and D. Moldovan (2005). Exploiting ontologies for automatic image annotation. In *Proceedings of the international ACM SIGIR conference on Research and development in information retrieval*, pp. 552–558.
- Stanek, M. and O. Maier (2010). The Wroclaw University of Technology Participation at ImageCLEF 2010 Photo Annotation Track. In *Working Notes of CLEF 2010, Padova, Italy*.
- Strube, M. and S. P. Ponzetto (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 1419–1424. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.
- Struyf, J., S. Džeroski, H. Blockeel, and A. Clare (2005). Hierarchical multi-classification with predictive clustering trees in functional genomics. In *Progress in Artificial Intelligence*, Volume 3808, pp. 272–283. LNCS.
- Su, J.-H., C.-L. Chou, C.-Y. Lin, and V. S. Tseng (2010). Effective image semantic annotation by discovering visual-concept associations from image-concept distribution model. In *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 42–47.
- Sun, A. and E. P. Lim (2001). Hierarchical text classification and evaluation. In *Proceedings of the IEEE International Conference on Data Mining*, Volume 528. California, USA.
- Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management* 28(4), 467–490.
- Tague-Sutcliffe, J. and J. Blustein (1995). A statistical analysis of the TREC-3 data. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pp. 385–398.
- Tahir, M. A., Y. Fei, M. Barnard, M. Awais, K. Mikolajczyk, and J. Kittler (2010). The University of Surrey Visual Concept Detection System at ImageCLEF 2010: Working Notes. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey.
- Thom, J. A. and F. Scholer (2007). A comparison of evaluation measures given how users perform on search tasks. In *Australasian Document Computing Symposium*, pp. 100–103.
- Tieu, K. and P. Viola (2000). Boosting image retrieval. In *Proceedings of international Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 228–235. Published by the IEEE Computer Society.
- Tommasi, T. and T. Deselaers (2010). The Medical Image Classification Task. In H. Müller, P. Clough, T. Deselaers, and B. Caputo (Eds.), *ImageCLEF - Experimental Evaluation of Visual Information Retrieval*, Chapter 12, pp. 221–238. Springer.
- Torralba, A. and A. A. Efros (2011). Unbiased Look at Dataset Bias. In *To appear in: Proceedings of the international Conference on Computer Vision and Pattern Recognition (CVPR)*.

- Tsai, C. F. and C. Hung (2008). Automatically annotating images with keywords: A review of image annotation systems. *Recent Patents on Computer Science* 1(1), 55–68.
- Tsoumakas, G., I. Katakis, and I. Vlahavas (2010). Mining Multi-label Data. In O. Maimon and L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook*, pp. 667–685. Springer.
- Tsoumakas, G. and I. Vlahavas (2007). Random k-labelsets: An ensemble method for multilabel classification. In *Proceedings of European Conference on Machine Learning*.
- Ulges, A., C. Schulze, D. Keysers, and T. Breuel (2008). Identifying relevant frames in weakly labeled videos for training concept detectors. In *Proceedings of the International Conference on Content-based Image and Video Retrieval*, pp. 9–16.
- Vailaya, A., M. A. T. Figueiredo, A. Jain, and H. Zhang (1999). Content-Based Hierarchical Classification of Vacation Images. In *Proceedings of the International Conference on Multimedia Computing and Systems*, pp. 518–523.
- van de Sande, K. E. A. and T. Gevers (2010). The University of Amsterdam’s Concept Detection System at ImageCLEF 2010. In *Working Notes of CLEF 2010, Padova, Italy*.
- van de Sande, K. E. A., T. Gevers, and A. W. M. Smeulders (2010). The University of Amsterdam’s Concept Detection System at ImageCLEF 2009. In C. Peters, T. Tsirikika, H. Müller, J. Kalpathy-Cramer, J. F. G. Jones, J. Gonzalo, and B. Caputo (Eds.), *Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross-Language Evaluation Forum 2009, Revised Selected Papers*, Corfu, Greece. LNCS.
- van de Sande, K. E. A., T. Gevers, and C. G. M. Snoek (2010). Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(9), 1582–1596.
- Van Gemert, J. C., C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek (2010). Visual Word Ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(7), 1271–1283.
- van Rijsbergen, C. J. (1979). Evaluation. In *Information Retrieval* (2nd ed.), Chapter 7, pp. 112–140. Butterworth-Heinemann Ltd.
- Verbeek, J., M. Guillaumin, T. Mensink, and C. Schmid (2010). Image annotation with Tagprop on the MIR-Flickr set. In *Proceedings of the International Conference on Multimedia Information Retrieval (MIR)*, pp. 537–546.
- Véronis, J. (1998). A study of polysemy judgements and inter-annotator agreement. In *Programme and advanced papers of the Senseval workshop*.
- Verspoor, K., J. Cohn, S. Mniszewski, and C. Joslyn (2006). A categorization approach to automated ontological function annotation. *Protein Science* 15(6), 1544–1549.
- Vijayanarasimhan, S. and K. Grauman (2009). What’s it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2262–2269.
- Volkmer, T., J. A. Thom, and S. M. M. Tahaghoghi (2007). Modeling human judgment of digital imagery for multimedia retrieval. *IEEE Transactions on Multimedia* 9(5), 967–974.

- von Ahn, L. and L. Dabbish (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 319–326.
- Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management* 36(5), 697–716.
- Voorhees, E. M. (2002). The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems*, pp. 143–170. Springer.
- Voorhees, E. M. and C. Buckley (2002). The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 11–15.
- Voorhees, E. M. and D. K. Harman (2005). *TREC: Experiment and evaluation in information retrieval*. MIT Press.
- Wang, Y., T. Mei, S. Gong, and X. S. Hua (2009). Combining global, regional and contextual features for automatic image annotation. *Pattern Recognition* 42(2), 259–266.
- Webber, W., A. Moffat, and J. Zobel (2007). Score standardization for robust comparison of retrieval systems. In *Proceedings of 12th Australasian Document Computing Symposium*, pp. 1–8.
- Weber, M., M. Welling, and P. Perona (2000). Towards automatic discovery of object categories. In *Proceedings of international Conference on Computer Vision and Pattern Recognition (CVPR)*, Volume 2, pp. 101–108.
- Weinberger, K. Q., M. Slaney, and R. van Zwol (2008). Resolving tag ambiguity. In *Proceedings of the International Conference on Multimedia*, pp. 111–120.
- Welinder, P. and P. Perona (2010). Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 25–32.
- Wnuk, K. and S. Soatto (2008). Filtering Internet Image Search Results Towards Keyword Based Category Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8.
- Wu, F., J. Zhang, and V. Honavar (2005). Learning classifiers using hierarchically structured class taxonomies. In *Abstraction, Reformulation and Approximation*, Volume 3607, pp. 313–320. LNCS.
- Wu, L., X. S. Hua, N. Yu, W. Y. Ma, and S. Li (2008). Flickr distance. In *Proceedings of the sixteen ACM International conference on Multimedia*.
- Wu, S. and F. Crestani (2003). Methods for ranking information retrieval systems without relevance judgments. In *Proceedings of the 2003 ACM symposium on Applied computing*, pp. 811–816.
- Wu, Z. and M. Palmer (1994). Verb Semantics And Lexical Selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*.
- Xiao, J., J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba (2010). SUN Database: Large-scale Scene Recognition from Abbey to Zoo. In *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*.

- Xie, L., A. Natsev, M. Hill, J. Smith, and A. Phillips (2010). The accuracy and value of machine-generated image tags: design and user evaluation of an end-to-end image tagging system. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR)*, pp. 58–65.
- Xu, H., X. Zhou, M. Wang, Y. Xiang, and B. Shi (2009). Exploring Flickr’s related tags for semantic annotation of web images. In *Proceeding of the ACM International Conference on Image and Video Retrieval (CIVR)*.
- Yan, T., V. Kumar, and D. Ganesan (2010). CrowdSearch: exploiting crowds for accurate real-time image search on mobile phones. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pp. 77–90.
- Yavlinsky, A., E. Schofield, and S. Rüger (2005). Automated image annotation using global features and robust nonparametric density estimation. In *Image and Video Retrieval*, Volume 3568, pp. 507–517. LNCS.
- Yilmaz, E. and J. A. Aslam (2006). Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 111.
- Yilmaz, E., J. A. Aslam, and S. Robertson (2008). A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 587–594.
- Yilmaz, E., E. Kanoulas, and J. A. Aslam (2008). A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 603–610.
- Yuen, J., B. Russell, C. Liu, and A. Torralba (2009). LabelMe video : Building a Video Database with Human Annotations. In *International Conference on Computer Vision*.
- Yuen, M. C., L. J. Chen, and I. King (2009). A survey of human computation systems. In *International Conference on Computational Science and Engineering*, pp. 723–728.
- Zesch, T. and I. Gurevych (2010). Wisdom of crowds versus wisdom of linguists—measuring the semantic relatedness of words. *Natural Language Engineering* 16(01), 25–59.
- Zhang, D., B. Liu, C. Sun, and X. Wang (2010). Random Sampling Image to Class Distance for Photo Annotation. In *Working Notes of CLEF 2010, Padova, Italy*.
- Zhang, M. L. and Z. H. Zhou (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition* 40(7), 2038–2048.
- Zhao, R. and W. I. Grosky (2000). From Features to Semantics : Some Preliminary Results. In *IEEE International Conference on Multimedia and Expo(ICME)*, pp. 679–682.
- Zobel, J. (1998). How Reliable are the Results of Large-Scale Information Retrieval Experiments? In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 307–314.

Own publications

- Brandenburg, K., C. Dittmar, M. Gruhne, J. Abeßer, H. Lukashevich, P. Dunker, D. Gärtner, K. Wolter, S. Nowak, and H. Grossmann (2009). Music Search and Recommendation. In B. Furht (Ed.), *Handbook of Multimedia for Digital Entertainment and Arts*, Springer.
- Dunker, P., C. Dittmar, A. Begau, S. Nowak, and M. Gruhne (2009). Semantic High-Level Features for Automated Cross-Modal Slideshow Generation. In *International Workshop on Content-Based Multimedia Indexing (CBMI)*.
- Dunker, P., S. Nowak, A. Begau, and C. Lanz (2008). Content-based Mood Classification for Photos and Music: A generic multi-modal Classification Framework and Evaluation Approach. In *Proceedings of the 1st International Conference on Multimedia Information Retrieval (ACM MIR), Vancouver, Canada*.
- Dunker, P., R. Paduschek, C. Dittmar, S. Nowak, and M. Gruhne (2009). Evaluation of an Image and Music Indexing Prototype. In *Workshop Audiovisuelle Medien, WAM*.
- Grubinger, M., S. Nowak, and P. Clough (2010). Data Sets created in ImageCLEF. In H. Müller, P. Clough, T. Deselaers, and B. Caputo (Eds.), *ImageCLEF - Experimental Evaluation of Visual Information Retrieval*, The Information Retrieval Series, Chapter 2, pp. 19–43. Springer.
- Lanz, C., S. Nowak, and U. Kühhirt (2010). Determination of Categories for Tagging and Automated Classification of Film Scenes. In *EuroITV 2010, 8th European Conference on Interactive TV and Video*.
- Lanz, C., S. Nowak, and H. Lukashevich (2010). Automated Classification of Film Scenes based on Film Grammar. In *Workshop Audiovisuelle Medien, WAM*.
- Lukashevich, H., S. Nowak, and P. Dunker (2009). Using One-Class SVM Outliers Detection for Verification of Collaboratively Tagged Image Training Sets. In *IEEE International Conference on Multimedia and Expo (ICME)*.
- Nagel, K., S. Nowak, U. Kühhirt, and K. Wolter (2011). The Fraunhofer IDMT at ImageCLEF 2011 Photo Annotation Task. In *Working Notes of CLEF 2011, Amsterdam, The Netherlands*.
- Nowak, S. (2010a). ImageCLEF@ICPR Contest: Challenges, Methodologies and Results of the Photo Annotation Task. In *20th International Conference on Pattern Recognition, ICPR*.
- Nowak, S. (2010b). Overview of the Photo Annotation Task in ImageCLEF@ICPR. In D. Unay, S. Aksoy, and Z. Cataltepe (Eds.), *Proceedings of the ICPR 2010 Contests*. LNCS.
- Nowak, S., C. Bastuck, and C. Dittmar (2008). Exploring Music Collections through Automatic Similarity Visualization. In *Tagungsband der DAGA Fortschritte der Akustik*, Dresden, Germany.
- Nowak, S. and P. Dunker (2009a). A Consumer Photo Tagging Ontology: Concepts and Annotations. In *THESEUS/ImageCLEF Pre-Workshop, Corfu, Greece*.

- Nowak, S. and P. Dunker (2009b). Overview of the CLEF 2009 Large-Scale Visual Concept Detection and Annotation Task. In *CLEF working notes 2009, Corfu, Greece*.
- Nowak, S. and P. Dunker (2010). Overview of the CLEF 2009 Large-Scale Visual Concept Detection and Annotation Task. In C. Peters, T. Tsirikika, H. Müller, J. Kalpathy-Cramer, J. F. G. Jones, J. Gonzalo, and B. Caputo (Eds.), *Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross-Language Evaluation Forum (CLEF 2009), Revised Selected Papers*, Corfu, Greece. LNCS.
- Nowak, S., P. Dunker, and R. Paduschek (2008). A generic Framework for the Evaluation of content-based Image and Video Analysis Tasks in the Core Technology Cluster of THESEUS. Quaero/ImageCLEF Preworkshop, Aarhus, Denmark.
- Nowak, S., P. Dunker, and R. Paduschek (2009, September). THESEUS Meets ImageCLEF: Combining Evaluation Strategies for a New Visual Concept Detection Task 2009. In C. Peters, D. Giampiccol, N. Ferro, V. Petras, J. Gonzalo, A. Peñas, T. Deselaers, T. Mandl, G. Jones, and M. Kurimo (Eds.), *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, LNCS, Aarhus, Denmark.
- Nowak, S., A. Hanbury, and T. Deselaers (2010). Object and Concept Recognition for Image Retrieval. In H. Müller, P. Clough, T. Deselaers, and B. Caputo (Eds.), *ImageCLEF - Experimental Evaluation of Visual Information Retrieval*, The Information Retrieval Series, Chapter 11, pp. 199–219. Springer.
- Nowak, S. and M. Huiskes (2010). New Strategies for Image Annotation: Overview of the Photo Annotation Task at ImageCLEF 2010. In *Notebook Papers, CLEF 2010 LABs and Workshop, 22-23 September, Padua Italy*.
- Nowak, S., A. Llorente, E. Motta, and S. Rüger (2010). The Effect of Semantic Relatedness Measures on Multi-label Classification Evaluation. In *ACM International Conference on Image and Video Retrieval (CIVR)*.
- Nowak, S. and H. Lukashevich (2009). Multilabel Classification Evaluation using Ontology Information. In *The 1st Workshop on Inductive Reasoning and Machine Learning on the Semantic Web -IRMLeS 2009, co-located with the 6th Annual European Semantic Web Conference (ESWC), Heraklion, Greece*.
- Nowak, S., H. Lukashevich, P. Dunker, and S. Rüger (2010). Performance Measures for Multilabel Evaluation: a Case Study in the Area of Image Classification. In *Proceedings of the international conference on Multimedia information retrieval (MIR)*, pp. 35–44.
- Nowak, S., K. Nagel, and J. Liebetrau (2011). The CLEF 2011 Photo Annotation and Concept-based Retrieval Tasks. In *CLEF 2011 working notes, Amsterdam, The Netherlands*.
- Nowak, S., R. Paduschek, and U. Kühhirt (2011). Photo Summary: Automated Selection of Representative Photos from a Digital Collection. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*.
- Nowak, S. and S. Rüger (2010). How reliable are Annotations via Crowdsourcing? A Study about Inter-annotator Agreement for Multi-label Image Annotation. In *ACM SIGMM International Conference on Multimedia Information Retrieval (ACM MIR), Philadelphia, Pennsylvania*.
- Paduschek, R., S. Nowak, and U. Kühhirt (2011). Automated Detection of Errors and Quality Issues in Audio-Visual Content. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*.

Schwarze, T., T. Riegel, S. Han, A. Hutter, S. Nowak, S. Ebel, C. Petersohn, and P. Ndjiki-Nya (2011). Role-based Identity Recognition for TV Broadcasts. *Multimedia Tools and Applications*, 1–20.