applied sciences

MDPI

Article

Image Analysis Using Human Body Geometry and Size Proportion Science for Action Classification

Syed Muhammad Saqlain [1], Anwar Ghani [1,*], Imran Khan [1], Shahbaz Ahmed Khan Ghayyur [1], Shahaboddin Shamshirband [2,3,*], Narjes Nabipour [4] and Manouchehr Shokri [5]

1 Department of CS & SE, International Islamic University, Islamabad 44000, Pakistan; syed.saqlain@iiu.edu.pk (S.M.S.); imran.khan@iiu.edu.pk (I.K.); shahbaz.ahmed@iiu.edu.pk (S.A.K.G.)
2 Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh, Viet Nam
3 Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh, Viet Nam
4 Institute of Research and Development, Duy Tan University, Da Nang 50000, Viet Nam; narjesnabipour@duytan.edu.vn
5 Faculty of civil engineering, Institute of Structural Mechanics (ISM), Bauhaus-Universität Weimar, 99423 Weimar, Germany
* Correspondence: anwar.ghani@iiu.edu.pk (A.G.); shahaboddin.shamshirband@tdtu.edu.vn (S.S.)

Received: 29 March 2020; Accepted: 24 April 2020; Published: 7 August 2020

Featured Application: The proposed technique is an application of human behavior analysis, analyzing six human behaviors. It may be applied in surveillance systems for abnormal events and action detection. Furthermore, the extended version of the application may be used in the context of the medical domain for automated patient care systems.

Abstract: Gestures are one of the basic modes of human communication and are usually used to represent different actions. Automatic recognition of these actions forms the basis for solving more complex problems like human behavior analysis, video surveillance, event detection, and sign language recognition, etc. Action recognition from images is a challenging task as the key information like temporal data, object trajectory, and optical flow are not available in still images. While measuring the size of different regions of the human body i.e., step size, arms span, length of the arm, forearm, and hand, etc., provides valuable clues for identification of the human actions. In this article, a framework for classification of the human actions is presented where humans are detected and localized through faster region-convolutional neural networks followed by morphological image processing techniques. Furthermore, geometric features from human blob are extracted and incorporated into the classification rules for the six human actions i.e., standing, walking, single-hand side wave, single-hand top wave, both hands side wave, and both hands top wave. The performance of the proposed technique has been evaluated using precision, recall, omission error, and commission error. The proposed technique has been comparatively analyzed in terms of overall accuracy with existing approaches showing that it performs well in contrast to its counterparts.

Keywords: action recognition; rule based classification; human body proportions; human blob

1. Introduction

Images are an important source of information sharing and have been used for many decades to represent actions, events, things, and scenes, etc. It is generally believed that an image speaks a thousand words and has served through its wide use in newspapers, posters, magazines, and books. Images containing different actions are easily understood by humans. Automatic image recognition is

an important area of research and a great deal of time and effort has been invested to achieve this goal. Human actions are a series of gestures generated by the human body, and their recognition includes analyzing them by matching the specified patterns. The recognition of human actions is widely used in solving different research problems like surveillance, activity analysis, and event recognition, etc. Over the last two decades, a lot of efforts have been made to recognize human actions from video content that has sufficient temporal and spatial information. However, a little work has been done for the same purpose using still images due to its challenging nature as static images have less information than video sequences.

Human detection is the first step in the recognition of human actions. Many well-known efforts are made for human detection using a histogram of oriented gradients [1–3]. However, in the last decade, it has been observed that the deep learning-based classifier has been performing better than the conventional feature extraction based classification. Ren et al. [4] used a faster regional-convolutional neural network (Faster R-CNN) to detect and locate the objects. This may be applied for the detection of a specific object like human detection and localization. Much of the work carried out in the field of action recognition includes silhouette-based features through depth images or vision-based generic features for 2D images with inherent limitations. Geometric features can be a good source of information used to solve the classification problem [5] and domain-specific features may perform better than the generic nature of global or local features [6].

This article presents a technique to address the problem of recognizing human actions from still images. The task of human detection and localization is accomplished through the use of faster R-CNN followed by background modeling (BM) and segmentation algorithm (SA). Geometric features have been extracted from the human blob and adopted in the context of human body size proportions science. The computed features are used to model the classification rules for six human actions: single-hand side wave, single-hand top wave, walking, standing, both hands side wave, and both hands top wave.

The rest of the paper has following organization: Section 2 presents the existing research, Section 3 presents the proposed research, Section 4 presents results and discussion over the produced results, while Section 5 concludes the presented research.

2. Literature Review

This section presents a review of the state of the art techniques relevant to the proposed research. It is further divided into four subsections: i.e., (1) Features based Action Recognition from 2D Videos, (2) Deep learning-based Action Recognition from 2D Videos, (3) Action Recognition using depth Videos, and (4) Action Recognition from still Images.

2.1. Features Based Action Recognition from 2D Videos

Evangelidis et al. [7] proposed local features descriptor from the human skeleton by generating view independent features covering 3D views. A Gaussian mixture model was used for generating Fisher kernels from skeletons which have been used as discriminant features. The action classification task was achieved through a linear support vector machine (SVM). Zhang et al. [8] proposed a methodology for the recognition of human actions. They formulated a global feature descriptor based on local features. All the local features of human body parts were calculated independently of overall human body actions. The local features were used for the recognition of global actions. Marín-Jiménez et al. [9] proposed a multiscale descriptor for human action recognition obtained through pyramids of optical flow histograms and tested their technique over standard datasets. Al-Ali et al. [10] explained contour and silhouette based approaches for human action classification in their book chapter. While investigating the contour-based techniques, they put the emphasis on four features i.e., chord-length, Cartesian coordinate, centroid-distance, and Fourier descriptors' features. On the other hand, for silhouette, they discussed a histogram of oriented optical flow, structural similarity index measure, and a histogram of oriented gradients. They tested the features through SVM and K-nearest neighbor

(KNN) classifiers. Veenendaal et al. [11] proposed their technique for the classification of human activity by extracting the human shapes from a sequence of frames and then used eigen and canonical space transformations to obtain binary state. After downsampling all the activity frames to a single frame, they classified it through decision rules. Wu et al. [12] represented human actions in the form of graphs and computed context-oriented graph similarities. The graph kernels were combined and used to train the classifiers.The local features used initially for representing the graph vertices and edges were the relationships between features in inter and intra frames. Veenendaal et al. [13] used a dynamic probabilistic network (DPN) for the classification of four human actions i.e., walking, object lifting, standing, and sitting. All the actions were captured from an indoor environment. Initially, they extracted the features through key regions i.e., legs, body, arms, and face and then temporal based links between these key regions were extracted. The dynamic links were then used as input to DPN that classified them as valid human actions. Abdulmunem et al. [14] used a combination of local and global descriptors through SVM for human action recognition. For representing the local descriptor, 3D-scale invariant feature transform (SIFT) features were used while for the global ones they used a histogram of oriented optical flow. The computational complexity is avoided by detecting the salient objects from the frames and only those frames were processed where objects were found. The authors validated their technique by performing experiments over standard datasets. The real-time human actions from videos are recognized by Liang et al. [15] focusing on the lower human limb based actions by detecting the hip joint as a first step. The motion information was gathered through the y-axis of the hip joint along with its acceleration and velocity. The motion information was subject to filtration through Kalman and wavelet transform. Human actions were defined through filtered information and classified through the dynamic Bayesian network. Luvizon et al. [16] used temporal and spacial local features calculated from a sequence of skeletons of humans taken from depth images. The authors used the KNN model for classification. A feature extractor from the skeleton of human images was presented [17] that could be used for classification purposes. It was tested on multiple datasets along with a user-generated dataset. View invariant features were extracted by Chou et al. [18] using a holistic set of features. Gaussian mixture model and nearest neighborhood were used for classification purposes. Human body parts based features, for twelve body parts, were exploited to represent different actions performed by a human [19]. The features were fed to the artificial neural networks (ANN) for classification and validated over KTH and Weizmann datasets. 3D Spatio-temporal gradient histograms were used to form a feature vector for action recognition in [20]. The gradients were supposed to work in arbitrary scales and parameter optimization regarding the action classification was evaluated as well. Interest points-based spatiotemporally windowed data [21] features were employed for human behavior classification while support vector machine-based human skeleton features [22] were presented for the same task as well. Multiclass support vector machines were used by Sharif et al. [23] extracting three types of feature vectors from the input frames i.e., local binary patterns, the histogram of oriented gradients, and Harlick features. The features selection was performed through Euclidean distance and joint entropy-PCA-based method. Finally, features were fed to the classifier for classification purposes. Another research work [24] used features from human skeletal and classification was achieved through kernel-based SVM.

2.2. Deep Learning Based Action Recognition from 2D Videos

Recently, deep learning has been a focused area of research [25]. Wang et al. [26] employed deep learning for action recognition from videos. They used convolutional neural networks that have been widely used for images. Zhu et al. [27] used the co-occurrence of features for joints of human skeletons. They used deep learning in a recurrent neural network for training using long short-term memory. Chéron et al. [28] worked on action classification through the convolutional neural network where feature representation was derived from a human pose. The pose descriptor combined motion and appearance information along with trajectories of body parts. Authors achieved better results than the state-of-the-art techniques. Pan et al. [29] defined the convolutional neural network models as

double deep as they can be composed in temporal and spatial layers. Authors argued that these models are suitable in scenarios where training data are very limited or the target concept is very complex. Kataoka et al. [30] defined transitional actions as ones that are in between the two action classes. Those were the states where an actor was transiting from one action to another. As there is not a huge difference between actions and transitional actions, in order to distinguish between both, they used convolutional neural network-based subtle motion descriptors. Once the actions and transitional actions are correctly classified, the next actions can be anticipated using a combination of them.

2.3. Action Recognition Using Depth Videos

Chen et al. [31] used depth videos for recognizing human actions. They generated depth decision maps from the front, top, and side views and motion information is measured from depth maps using local binary patterns (LBP) using two types of fusion i.e., (1) fusion of LBP features obtained from all the three views and (2) fusion of classification outcomes. They obtained better results than the state-of-the-art techniques. Chen et al. [32] used the Kinect depth sensor in combination with wearable inertial sensors. They obtained the data in three forms i.e., (1) joint positions of human skeletons, (2) depth images, and (3) signals from wearable inertial sensors. The output of the individual classifier for each of the input data are fused to classify the human action. The authors revealed that the result of fusing the outcome of three collaborative classifiers is better than the use of individual data separately. Li et al. [33] worked on recognizing the actions of humans from depth camera data arguing that a good technique must divide the human body into different parts and features must be extracted from each part. They further described that the combinations of feature descriptors should be of good discriminative nature but at the same time. Authors presented their technique which used part based features along with depth data by applying sparse based learning methods, consequently, producing reasonably better results. Jiang et al. [34] analyzed the contribution made by each human skeleton joint for different actions. The authors worked with 3D human skeletons achieved through the Kinect devices. Human joints have been used to form a feature vector for action recognition by Chaaraoui et al. [35] from RGB-D images. The RGB-D device produced 3D locations for the body joints which were later used for classification. Geodesic distances have been used by Kim et al. [36] to estimate human joints from image data collected by 3D depth sensors. The joints were calculated for body parts involving motion and the computed features were used in conjunction with SVM to classify the actions.

2.4. Action Recognition from Still Images

Chaaraoui et al. [37] presented their methodology where human actions were recognized using contour points of silhouette and learned through multi-view poses. They not only achieved a better computational complexity for real-time processing but variations in actions by different actors were handled as well. Guo and Lai [38] argued that human action recognition from images is unlike videos, as there is no temporal information available in still images. They discussed the state of the art techniques for action recognition from still images by providing a detailed survey and concluded by providing their views over those techniques. Zhao et al. [39] performed human action classification from still images exploiting the concept that the human has some periodic and symmetric pairs and their detection helps to identify discriminative regions for action classification. The authors evaluated their technique over four datasets. Sharma et al. [40] proposed methodology which recognized the human attributes along with actions in still images. To achieve their task, they identified the human body parts using the collection of templates. After localizing the required human body parts, the required attributes and actions were classified. Vishwakarma and Kapoor [41] used the human silhouette to recognize the action by extracting features from grids and cells of fix sizes.

3. Proposed Solution

The proposed methodology recognizes human action from the given input image. Initially, human detection and localization are achieved through the use of faster R-CNN [4] and post-processing analysis. Next, the task is to compute geometric features and afterward, classification rules using these geometric features are presented. The graphical representation of the proposed model is shown in Figure 1.

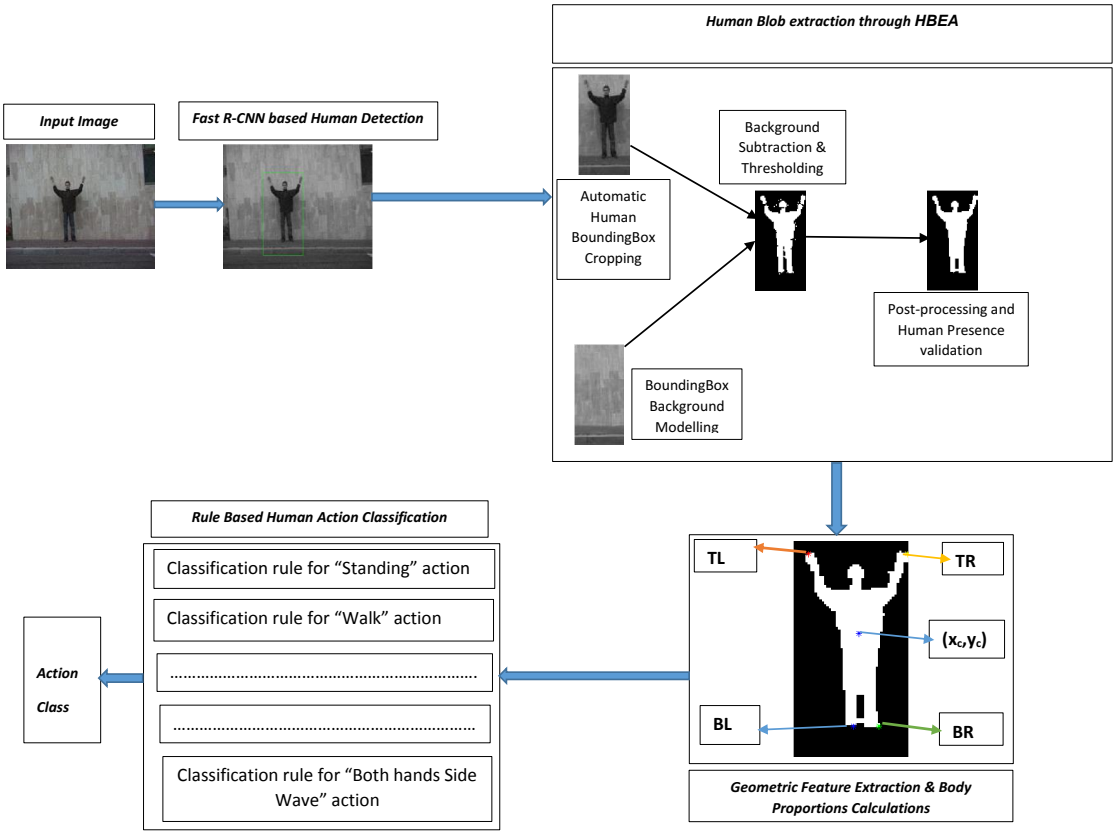


Figure 1. Workflow for the proposed technique.

3.1. Human Detection

Given an Image, I, having O_j objects: $1 \leq j \leq k$, we need to detect and localize human from it. To detect human in an image, the proposed technique uses faster R-CNN [4]. The general architecture of faster R-CNN has been presented in Figure 2.

It may be observed that the input image is provided to the convolutional layer that produces a convolutional feature map. Instead of using a selective search algorithm on the feature map for identification of the region proposals, a separate network is used for predicting them. The predicted region proposals are then reshaped using a region of interest (RoI) pooling layer that is ultimately used to classify the image within the proposed region and predict the offset values for the bounding boxes as well.

To deal with different scales and aspect ratios of human, anchors are used in the region proposal network (RPN). Each anchor is associated with a scale and an aspect ratio. Following the default setting of [4], three scales (1282, 2562, and 5122 pixels) and three aspect ratios (1:1, 1:2, and 2:1) have been used leading to $k = 9$ anchors at each location. Each proposal is parameterized relative to an anchor. Therefore, for a convolutional feature map of size $W \times H$, there are at most WHk possible proposals.

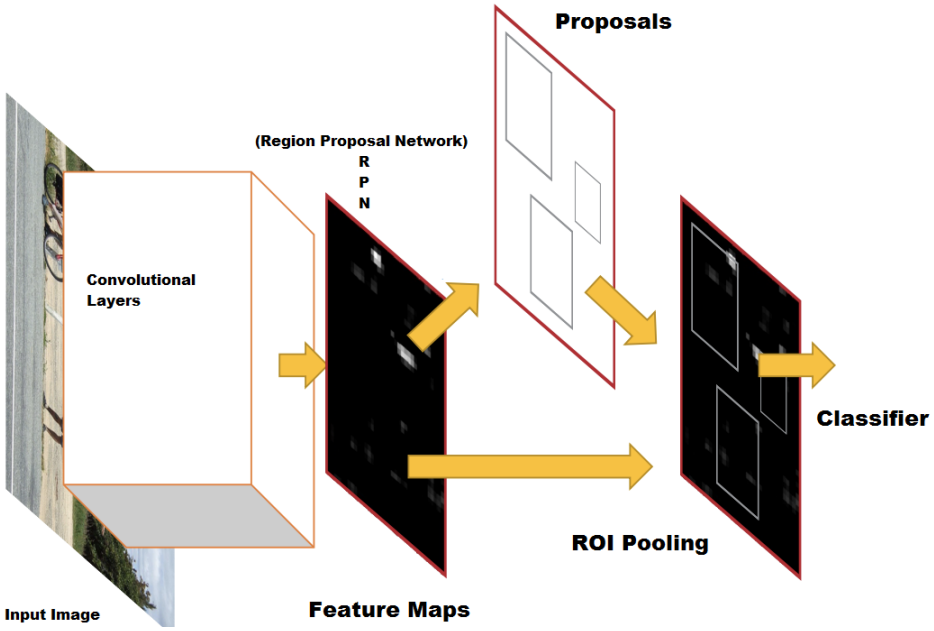


Figure 2. Faster-CNN architecture.

3.2. Human Segmentation

Bounding Box returned through faster R-CNN is a Quad Tuple i.e., $BB = \{X_{init}, Y_{init}, Width\_BB, Height\_BB\}$. Bounding Box is cropped automatically returning four points $\{(X_{init}, Y_{init}),\ (X_{init} + Height\_BB, Y_{init}),(X_{init}, Y_{init} + Width\_BB),(X_{init} + Height\_BB, Y_{init} + Width\_BB)\}$ termed as BBI and its dimensions are $Width\_BB \times Height\_BB$. Pixels of OI (Original Image) within the set of points i.r., $\{(X_{init}, Y_{init}),(X_{init} + Height\_BB, Y_{init}),(X_{init}, Y_{init} + Width\_BB),(X_{init} + Height\_BB, Y_{init} + Width\_BB)\}$, are marked as Black. Background of BBI is modelled through BM Algorithm 1.

Algorithm 1 Background Modeling of BBI Algorithm (BMA)

Input: Original Image,Bounding Box Image OI, BBI
Output: Background of BBI, BI

Create a new Image ,BI, of size $Width\_BB \times Height\_BB$ and initialize each of its pixel as black.
for row $r_i : X_{init} \leq i \leq X_{init} + Height\_BB$ **do**
  For each column $cj : Yinit \leq j \leq Y_{init} + \frac{Width\_BB}{2}$
  $LN = \{\frac{Width\_BB}{2}$ many left neighbours of $OI(r_i, c_j)\}$
  Find histogram of neighbouring pixels, $H = Histogram(LN)$;
  Find peak of Histogram, $P = Peak(H)$;
  Assign color value to pixels in Bounding Box of original image, $OI(r_i, c_j) = P$;
  Assign color to pixels of background image, $BI(r_i - X_{init} + 1, c_j - Y_{init} + 1) = P$;
end for
for row $r_i : X_{init} \leq i \leq X_{init} + Height\_BB$ **do**
  for column $c_j : Y_{init} + Width \geq j \geq Y_{init} + \frac{Width\_BB}{2}$ **do**
    $RN = \{\frac{Width\_BB}{2}$ many right neighbours of $OI(r_i, c_j)\}$
    Find histogram of neighbouring pixels, $H = Histogram(RN)$
    Find peak of Histogram, $P = Peak(H)$
    Assign color value to pixels in Bounding Box of original image, $OI(r_i, c_j) = P$
    Assign color to pixels of background image, $BI(r_i - X_{init} + 1, c_j - Y_{init} + \frac{Width\_BB}{2} - 1) = P$
  end for
end for
return BI

Next, the task is to segment human blob from the enclosing rectangle and is accomplished through presenting a segmentation algorithm (SA). Algorithm 2 takes the output of the BM algorithm, BI, along with BBI as its input and returns a segmented image, SI, having human blob (HB) as output.

Algorithm 2 Segmentation Algorithm (SA)

Input: BBI, BI
Output: Segmented Image, SI
1: For Segmented Image(SI), subtract background image(BI) from BBI. i.e., $SI = BBI - BG$
2: After applying thresholding over SI by producing binary level image.
3: Fill holes from blobs present in SI.
4: Apply dilation followed by erosion.
5: Apply Gaussian smoothing over SI resulted from above steps.
6: Retain blobs in SI having size greater than threshold.
7: Return SI having Human Blob(HB).

3.3. Feature Extraction

A segmented bounding box image (SI) obtained through the Segmentation Algorithm (SA) has both foreground (human blob) and background (black) pixels, we need to extract features from a human blob that would be used in classifying six human actions. SI is represented as $SI = \{FP \cup BP\}$, where $FP = \cup\{\forall(x_p, y_q) : SI(x_p, y_q)$ is foreground$\}$ and
$BP = \cup\{\forall(x_r, y_s) : SI(x_r, y_s)$ is background$\}$.

To deal with the human actions under the presented study, geometrical positions of the hands and feet are quite important. The graphical representation of geometrical features from the human blob is shown in Figure 3. In the case of a hand wave, the position of the hand is important, while, in the case of straight standing or walking step, the position of feet is important. These positions may be represented as discriminant features, but they need to be calculated with some reference point. We have defined the centroid of the human blob as a reference point for calculating the feature set. Centroid of the Human blob, (X_{cb}, Y_{cb}) is calculated using: $X_{cb} = \frac{1}{m}\Sigma_{i=1}^{m} x_i$ and $Y_{cb} = \frac{1}{n}\Sigma_{j=1}^{n} y_j$, where x_i and y_j represent pixel positions in FP. Boundary points of HB are extracted through finding the pixels p_j such that $p_j \in FP$ and in its 8-neighborhood at least one of the neighbor $n_p \in BP$. A boundary vector BV is created and all the boundary points are added to it. To obtain salient features, we divided the boundary vector (BV) of HB into four regions keeping the centroid of the HB as a reference point i.e., Top Left (TL), Top Right (TR), Bottom Left (BL), and Bottom Right (BR). The features for each of the actions are shown in Figure 4.

The idea behind dividing the HB into four regions lies in the physical positions of both the hands and feet i.e., hands are in the upper half region of HB, while the feet position of HB is in its lower half. In each of the four regions, the farthest points from the centroid are calculated as:

$$MDP_K = \max\{d_p = \sqrt{x_p - x_c)^2 + (y_p - y_c)^2)} : \forall p \in K\}, \text{where } K \in \{TL, TR, BL, BR\}. \quad (1)$$

The position of the farthest distant points in each region is an important clue in the recognition of human actions under study. Along with the position of the calculated point, the angular position of the point about the centroid of the HB is equally useful as well. Drawing a line from the distant point to the centroid would help in calculating the angular position of points i.e., angles, Θ_K, of each farther distant point with reference to centroid i.e.,

$$\Theta_K = \tan^{-1} \frac{(MDP_K(y) - y_c)}{(MDP_K(x) - x_c)} \quad (2)$$

It may be observed from Figure 4a–f that, for each of the six actions, distances between TL, TR, and BL, BR gives better clues for recognizing them. From Figure 4a, it is evident that the distance D_{TLR} is not high as TL and TR are close to head area, while from the Figure 4b it may be inspected that there

is a considerable value for D_{TLR} as the position of the right stretched hand is farther from the position of the head point, calculated as TR. The same can be established from Figure 4c–f. Distance between farthest points in the top and bottom regions are calculated using Euclidean distance as:

$$D_{TLR} = |MDP_{TL} - MDP_{TR}| = \sqrt{(MDP_{TL}(x) - MDP_{TR}(x))^2 + (MDP_{TL}(y) - MDP_{TR}(y))^2} \quad (3)$$

$$D_{BLR} = |MDP_{BL} - MDP_{BR}| = \sqrt{(MDP_{BL}(x) - MDP_{BR}(x))^2 + (MDP_{BL}(y) - MDP_{BR}(y))^2} \quad (4)$$

It is to be noted that feet are at the largest distance from the centroid of the body in the lower body region. D_{BLR} represents the distance between the extreme points in the lower part of the body i.e., the distance between feet. This distance gives a clue about whether the human under study is in standing or walking position. The ratio of D_{BLR} to the height of the blob gives step to height ratio (SHR) i.e.,

$$SHR = \frac{D_{BLR}}{HB\_Height} \quad (5)$$

Literature about physical dimensions of the human body reveals that there exist ratios between the size of different body parts to the height of human [42,43] i.e., length of the arm is approximately 0.44 of the height of the human i.e,

$$armLength = 0.44 * HB\_Height \quad (6)$$

Furthermore, the human body dynamics depict that the arm may be divided into two portions i.e., lower and upper arm. The proportion of the lower to the upper arm is nearly 4:3 and may be represented as:

$$upperarm = (\frac{3}{7}) * armLength \quad (7)$$

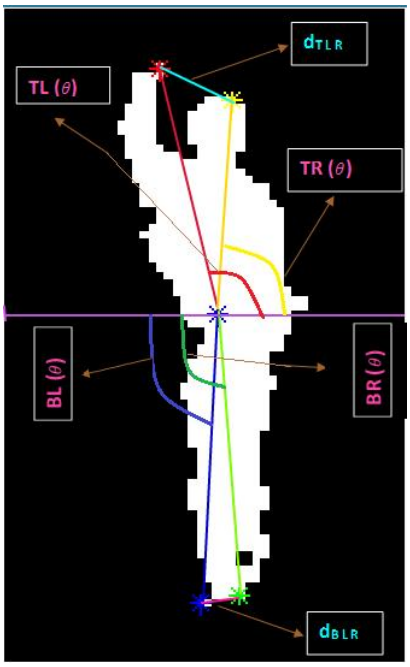
$$lowerarm = (\frac{4}{7}) * armLength \quad (8)$$

The length of the lower arm is the sum of the forearm and hand i.e.,

$$lowerarm = forearm + hand \quad (9)$$

The ratio between the length of the hand and the forearm is 2 : 3. We computed the length of the hand from the lower arm through the following relation i.e.,

$$handSize = (\frac{2}{5}) * lowerarm \quad (10)$$



8

Figure 3. Geometric feature extraction from human blob.

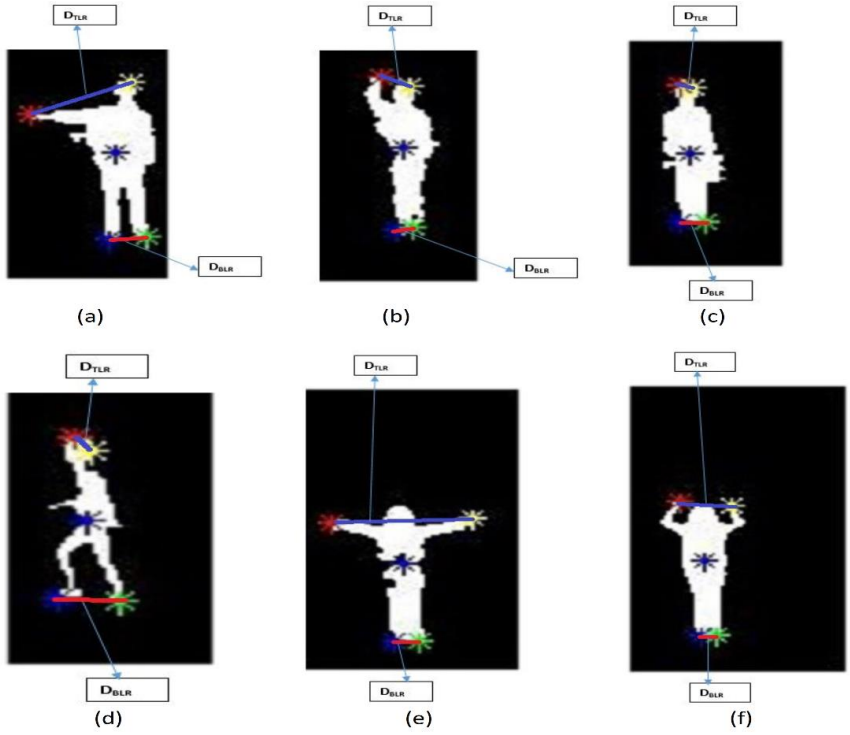


Figure 4. Representing geometric features for human actions. (**a**) single hand side wave; (**b**) single hand top wave; (**c**) standing; (**d**) Walking; (**e**) both hands side wave; (**f**) both hands top wave.

Fig. 2. Top and Bottom Distances for all the six actions

Algorithm 2: Human Height Calculation Algorithm (HHCA)

H **INPUT:** Foreground with human blob(FG),Human blob centroid ((X_{cb}, Y_{cb})), Bottom Left point(BL), Bottom Right point(BR)
OUTPUT: Human Blob Height (HB Height)

1: From the centroid, (X_{cb}, Y_{cb}), of FG trace upward on 90^{θ} till a background pixel is found.

2: Take position as (x_b,y_b) and as: Top_x=x-1; Top_y=y;

3: Mark (Top_x,Top_y) as center of head position.

4: Find mid column for bottom left and bottom right ($Bottom_y$)

5: From coordinates of BL and BR points find mid of column position i.e., $Bottom_y = \frac{(BL_y + BR_y)}{2}$

6: Find maximum row from BL and BR coordinate points i.e.,
$Bottom_{x=}max(BL_x, BR_x)$

3.4. Classification

Based on the extracted feature set, the rule-based classification is presented. The classification rules for all of the six actions are modeled and presented in the following subsection.

3.4.1. Case Standing

When a person is in a standing position, his hands are not in a stretched position. Either they are in parallel to the body in a downward direction or around the chest. In both of the cases, the extreme point TL and TR from the centroid are around the head area of the human. The width of the human head approximately matches the length of the hand. Thus, in a standing position, D_{TLR} would be lesser than or equal to the hand length of the human having $D_{TLR}(\theta)$ is not very large. From the reviews of the step science, it is observed that the ratio of the human step size to his height is in between 0.41 to 0.45. However, the human actions in still images, a person could be in a half step to full step in his walk. Through the experimental observations, it is deduced that a person in the standing position would have SHR lesser than 0.25. By combining all of the feature attributes, the rule for classifying a human in the standing action is presented as:

$$D_{TLR}(\theta) < 10 \text{ AND } D_{TLR} \leq handSize \text{ AND } SHR < 0.25 \tag{11}$$

3.4.2. Case Walking Step

In a posture of a walking step, the lower part of the human body shows significant changes. As mentioned in the standing case, the ration of SHR to the height of the human is 0.41 to 0.45. This is the ratio when a human is walking and having a full stretched step. However, the image may be one of a frame from all the sequences and the size of the step would not be accurate in the range of 0.41 to 0.45 of the human height. Furthermore, from most of the dataset images, the SHR varies from 0.25 to 0.42 for a human in walking posture. Angles from centroid point to extreme points in BL and BR regions are wider for walking step posture than that of standing action. By combining both SHR and $D_{BLR}(\theta)$, the discriminating rule for classifying a human in walking posture is presented as i.e.,

$$D_{BLR}(\theta) > 15 \text{ AND } SHR > 0.25 \tag{12}$$

3.4.3. Case Single Hand Side Wave

In case of recognizing the human action of waving a single hand on sidewise, it only needs to inspect the upper portion of the body i.e., TL and TR regions need to be focused. In the case of Right-hand wave, $TR(\theta)$ gets larger than 125 degrees relative to the centroid of the human blob, whilst the extreme point in the TR region remains near the head region having an angle closer to 90 degrees. The distance between extreme points of TL and TR regions is more than the length of the arm. The classification rule may be described as:

$$TL(\theta) > 125 \text{ AND } TR(\theta) > 75 \text{ AND } D_{TLR} > armLength. \tag{13}$$

Similarly for the left hand wave classification, angles get reversed as extreme TL point gets near head region:

$$TL(\theta) < 105 \text{ AND } TR(\theta) < 55 \text{ AND } D_{TLR} > armLength \tag{14}$$

3.4.4. Case Single Hand Top Wave

When a human waves his hand in top direction, the angle of the waving hand gets closer to the head position. As only one hand is waved, say the right one, an extreme TR point rests near the head position while the position of the hand can be quite closer or away from the head. When a waving hand has a closer angle with respect to the position of the head, the segmentation limitations might misclassify it to the standing posture. Likewise, when $TL(\theta)$ gets wider, then there may be some point where Top Wave and Side Wave have the same boundary point, and it may get misclassified. To tackle all of these issues, the proposed classification rule has used the combinational privilege of different features. The classification rule for the right-hand wave is given as:

$$TL(\theta) > 97.5 \text{ AND } TL(\theta) \leq 125) \text{ AND } TR(\theta) > 75 \text{ AND } D_{TLR} > hand\_Size \text{ AND } D_{TLR} < (\frac{2}{3}) * armLength \quad (15)$$

Just like the above, we can write the rule for the Left hand Top wave:

$$TR(\theta) < 82.5 \text{ AND } TL(\theta) \geq 55) \text{ AND } TL(\theta) < 105 \text{ AND } D_{TLR} > hand\_Size \text{ AND } D_{TLR} < (\frac{2}{3}) * armLength \quad (16)$$

3.4.5. Case Both Hands Side Wave

The proposed rule for Classifying both hands side wave is a combination of the left and right-hand side wave along with taking D_{TLR} into account. As defined by [44], the size of the Wingspan of a human is the same as his height. There are some cases as well when the direction of the wingspan is slightly upward resulting in reduced wingspan size. Thus, the proposed rule combines the mentioned constraints over the selected features and is defined as:

$$TL(\theta) > 125 \text{ AND } TR(theta) < 55 \text{ AND } D_{TLR} > 1.5 * armLength \quad (17)$$

3.4.6. Case Both Hands Top Wave

The proposed classification rule for both hands Top wave is a combination of single hand left and right top wave rules. The maximum distance between TL and TR extreme points must be at least the same as the length of the forearm and should not be greater than 1.5 of armLength. The minimum and maximum values for TL_θ and $TR(\theta)$ are also used as discriminating features. The proposed rule is described as:

$$\begin{gathered}(TL(\theta) > 97.5 \text{ AND } TL(\theta) \leq 125) \text{ AND} \\ (TR(\theta) < 82.5 \text{ AND } TR(theta) \geq 55)) \text{ AND} \\ (D_{TLR} > forearm \text{ AND } D_{TLR} \leq 1.5 * armLength))\end{gathered} \quad (18)$$

4. Results and Discussion

In this section, the details about dataset, evaluations metrics, and results achieved through the implementation of the proposed technique have been presented.

4.1. Dataset and System Platform

To evaluate the performance of the proposed methodology, various experiments are performed over the Weizmann dataset [45]. The dataset is used for six of the actions i.e., Standing, Walking, Single Hand Side Wave, Single Hand Top Wave, Both Hands Side Wave and Both Hands Top Wave. The action images are extracted from the videos having a human body in the motions of walking, jumping, bending, waving with one hand, and both hands. The six potential action classes have been presented in Table 1.

Table 1. Action Class labels with abbreviations.

	Action Class	Abbreviations
1	Single Hand Side Wave	SHSW
2	Single Hand Top Wave	SHTW
3	Standing	Standing
4	Walking	Walk
5	Both Hands Side Wave	BHSW
6	Both Hands Top Wave	BHTW

Images for all the six potential action classes are extracted from the videos of eight different actors, i.e., Darya, Denis, Eli, Ido, Ira, Lena, Lyova, and Moshe. Five of the actors depicting the actions were male while three were female. The human body dimensions of all the actors were different with respect to their heights, postures of walking, standing, and waving hands.

To conduct experiments, MATLAB 2015 (MathWorks, Natick, MA, USA) on a machine with the processing speed of 2.14 GHz Core i5 and 6 GB RAM has been used for implementing the proposed approach.

4.2. Performance Evaluation Metrics

The following evaluation metrics have been used to measure the performance of the proposed technique i.e., precision, recall, accuracy, F-score, omission error, and commission error. Each of the performance parameters has been briefly explained in the following subsections.

4.2.1. Precision

The precision score describes the ability of the classifier not to label a negative example as positive. The precision score can further be described as evaluating the probability that a positive prediction made by the classifying engine is in fact positive. The score ranges [0, 1], with 0 being the worst possible score and 1 being perfect. The Precision score is defined:

$$Precision = \frac{(\Sigma TruePositive)}{(\Sigma TruePositive + \Sigma FalsePositive)} \tag{19}$$

4.2.2. Omission Error

Given m many actions in class C_j, the Omission error represents actions that belong to C_j but were not accurately classified as being in the C_j class:

$$OmissionError = 1 - Precision \tag{20}$$

4.2.3. Recall

The Recall score describes the ability of a classifier not to identify a positive example as negative. The score ranges [0, 1], with 0 being the worst possible score and 1 being perfect. The Recall score can be further described as:

$$Recall = \frac{(\Sigma TruePositive)}{(\Sigma TruePositive + \Sigma FalseNegative)} \tag{21}$$

4.2.4. Commission Error

Given m many actions in a class C_j, the Commission error represents actions belonging to a different class but were inaccurately classified as being in the C_j class. Commission error is defined in relationship to Recall, as:

$$CommissionError = 1 - Recall \tag{22}$$

4.2.5. F-Score

F-Score is defined as a measure that provides a balance between recall and precision or it may be said as a harmonic mean of recall and precision. It may further be represented as:

$$F - Score = 2\frac{(Precision * Recall)}{(Precision + Recall)} \tag{23}$$

4.2.6. Accuracy

Accuracy is an important but simplistic measure of how often a classifier makes a correct prediction. It is depicted as the ratio between the number of correct predictions versus the total number of predictions. Overall accuracy represents the total classification accuracy:

$$Accuracy = \frac{(\Sigma TruePositive + \Sigma TrueNegative)}{(\Sigma TruePositive + \Sigma FalsePositive + \Sigma TrueNegative + \Sigma FalseNegative)} \tag{24}$$

4.3. Results

In this subsection, experimental results are presented when the proposed technique is applied on action image dataset. Figure 5 is showing the results through each step of proposed technique.

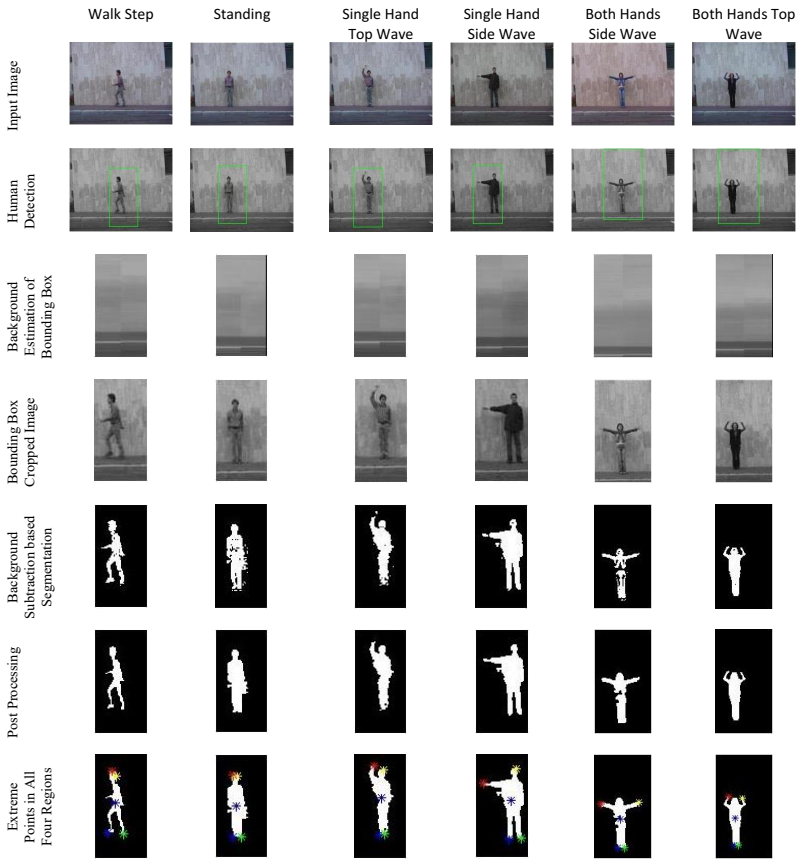


Figure 5. Results of each step through the proposed technique.

ndjndjm

Table 2 is showing a classification matrix, representing statistical results in the case when the proposed technique is applied over Daria images. Originally, 278 images of Daria having all the six actions were tested. The proposed technique achieved 100%, 96.3%, 83.1%, 86.9%, 91.2%, and 100% recall for SHSW, SHTW, walk, standing, BHSW, and BHTW actions, respectively, while there were

100% , 92.9% , 92.5%, 81.5%, 100%, and 93% precision values for them, respectively. An overall accuracy of 91.4% is achieved while, out of all the images, 9 remained un-classified.

Table 2. Classification matrix for Daria images.

	SHSW	*SHTW*	*Walk*	*Standing*	*BHSW*	*BHTW*	*Total Images Classified*	*Recall*	*Commission Error*	*Un-Classified Images*	*Total Images*	*Un-Classification (%)*
SHSW	**21**	0	0	0	0	0	21	100%	0%	3	24	12.5%
SHTW	0	**52**	0	2	0	0	54	96.3%	3.7%	0	54	0%
Walk	0	0	**49**	10	0	0	59	83.1%	16.9%	0	59	0%
Standing	0	4	4	**53**	0	0	61	86.9%	13.1%	2	63	3.2%
BHSW	0	0	0	0	**31**	3	34	91.2%	8.8%	4	38	10.5%
BHTW	0	0	0	0	0	**40**	40	100%	0%	0	40	0
Total Images	21	56	53	65	31	43	n = 269			9	m = 278	
Precision	100%	92.9%	92.5%	81.5%	100%	93.0%	Overall Accuracy(91.4%)					Overall Un-Classification(4.4%)
Ommission Error	0.0%	7.1%	7.5%	18.5%	0.0%	7.0%						

Table 3 is a classification matrix for the six defined actions achieved after implementing the proposed technique using the Denis images. A total of 188 images having all the six actions were tested and, out of those 188 images, 10 were un-classified while from the remaining 178 images 91.6% actions were accurately classified giving 100%, 95%, 81.9%, 91.7%, 100 %, and 100% recall and 100%, 82.6%, 100%, 80.0%, 100%, and 100% precision for SHSW, SHTW, walk, standing, BHSW, and BHTW actions, respectively.

Table 3. Classification matrix for Denis images.

	SHSW	*SHTW*	*Walk*	*Standing*	*BHSW*	*BHTW*	*Total Images Classified*	*Recall*	*Commission Error*	*Un-Classified Images*	*Total Images*	*Un-Classification (%)*
SHSW	**23**	0	0	0	0	0	23	100%	0%	2	25	8.0%
SHTW	0	**19**	0	1	0	0	20	95.0%	5.0%	2	22	9.1%
Walk	0	0	**45**	10	0	0	55	81.9%	18.1%	0	55	0%
Standing	0	4	0	**44**	0	0	48	91.7%	8.3%	0	48	0%
BHSW	0	0	0	0	**17**	0	17	100%	0%	3	20	15.0%
BHTW	0	0	0	0	0	**15**	15	100%	0%	3	18	16.7%
Total Images	23	23	45	55	17	15	n = 178			10	m = 188	Overall Un-Classification(5.3%)
Precision	100%	82.6%	100%	80.0%	100%	100%	Overall Accuracy(93.7%)					
Omission Error	0.0%	17.4%	0%	20.0%	0%	0%						

The classification matrix for Ira images is shown in Table 4. In this experiment, a total of 267 images are tested to evaluate the proposed technique. The unclassified images are 18 while the rest of the 249 images are classified. The overall accuracy for Ira actions is 93.2% while 93.9%, 92.9%, 86.2%, 100%, 90.9%, 90.9% recall and 91.2%, 95.1%, 100%, 87.8%, 100%, and 90.9% precision for SHSW, SHTW, walk, standing, BHSW and BHTW actions, respectively, are achieved.

Table 4. Classification matrix for Ira images.

	SHSW	*SHTW*	*Walk*	*Standing*	*BHSW*	*BHTW*	*Total Images Classified*	*Recall*	*Commission Error*	*Un-Classified*	*Total Images*	*Un-Classification (%)*
SHSW	**31**	2	0	0	0	0	33	93.9%	6.1%	1	34	2.9%
SHTW	3	**39**	0	0	0	0	42	92.9%	7.1%	7	49	14.3%
Walk	0	0	**50**	8	0	0	58	86.2%	13.8%	1	59	1.7%
Standing	0	0	0	**72**	0	0	72	100%	0%	3	75	4.0%
BHSW	0	0	0	0	**20**	2	22	90.9%	9.1%	4	26	15.4%
BHTW	0	0	0	2	0	**20**	22	90.9%	9.1%	2	24	8.3%
Total Images	34	41	50	82	20	22	n = 249			18	m = 267	Overall Un-Classification(6.7%)
Precision	91.2%	95.1%	100%	87.8%	100%	90.9%	Overall Accuracy(93.2%)					
Omission Error	8.8%	4.9%	0.0%	12.2%	0%	9.1%						

In Table 5, classification results for "Eli" have been presented showing precision, recall, overall accuracy, and other evaluation metrics. It may be observed that a total of 281 images of Eli are tested that contained all the six actions. A total of 273 images are classified leaving eight of them unclassified. Overall classification accuracy over Eli images is 91.7% while 8.3% actions are misclassified. There is 98.2%, 89.2%, 92.3%, 86%, 95.7%, and 85.2% recall, and 100%, 92.6%, 95.6%, 71.2%, 100%, and 95.8% precision for SHSW, SHTW, walk, standing, BHSW, and BHTW actions, respectively, is recorded. Least precision is for standing action while the least recall is for BHTW with an 85.2% score.

Table 5. Classification matrix for Eli images.

	SHSW	*SHTW*	*Walk*	*Standing*	*BHSW*	*BHTW*	*Total Images Classified*	*Recall*	*Commission Error*	*Un-Classified Images*	*Total Images*	*Un-Classification (%)*
SHSW	**56**	1	0	0	0	0	57	98.2%	1.8%	2	59	3.4%
SHTW	0	**50**	0	6	0	0	56	89.3%	10.7%	0	56	0.0%
Walk	0	0	**65**	5	0	0	70	92.3%	7.7%	0	70	0.0%
Standing	0	3	3	**37**	0	0	43	86.0%	14.0%	0	43	0.0%
BHSW	0	0	0	0	**22**	1	23	95.7%	4.3%	1	24	4.2%
BHTW	0	0	0	4	0	**23**	27	85.2%	14.8%	2	29	6.9%
Total Images	56	54	68	52	22	24	n = 276			5	m = 281	Overall Un-Classification(1.8%)
Precision	100%	92.6%	95.6%	71.2%	100%	95.8%	Overall Accuracy(91.7%)					
Omission Error	0	7.4%	4.4%	28.8%	0%	4.2%						

Table 6 is showing classification matrix results for the actions performed by "Ido". The original dataset contained 196 images having all six action images. Out of them, 10 images were not classified. For the images where the proposed technique is able to classify, the overall accuracy is 89.2%. It may be observed that the proposed technique achieved 88.5%, 80.8%, 85.4%, 93.5%, 96.4%, and 89.5% recall, and 100%, 77.8%, 100%, 78.6%, 100%, and 94.4% precision for the SHSW, SHTW, walk, standing, BHSW, and BHTW actions, respectively.

Table 6. Classification matrix for Ido images.

	SHSW	*SHTW*	*Walk*	*Standing*	*BHSW*	*BHTW*	*Total Images Classified*	*Recall*	*Commission Error*	*Un-Classified Images*	*Total Images*	*Un-Classification (%)*
SHSW	**23**	3	0	0	0	0	26	88.5%	11.5%	1	27	3.7%
SHTW	0	**21**	0	5	0	0	26	80.8%	19.2%	2	28	7.1%
Walk	0	0	**35**	6	0	0	41	85.4%	14.6%	2	43	4.7%
Standing	0	3	0	**43**	0	0	46	93.5%	6.5%	0	46	0.0%
BHSW	0	0	0	0	**27**	1	28	96.4%	3.6%	2	30	6.7%
BHTW	0	0	0	2	0	**17**	19	89.5%	10.5%	3	22	13.6%
Total Images	23	27	35	56	27	18	n = 186			10	m = 196	Overall Un-Classification(5.1%)
Precision	100%	77.8%	100%	76.8%	100%	94.4%	Overall Accuracy(89.2%)					
Omission Error	0%	22.2%	0%	23.2%	0%	5.6%						

The classification matrix for "Lena" images is presented in Table 7 obtained by testing the proposed technique over a total of 214 action images. The number of unclassified images is 9 while 205 have been successfully classified. The overall accuracy for "Lena" images is 93.7%. The matrix shows that the proposed technique achieved 100%, 91.3%, 97.1%, 85.7%, 91.7%, and 95.7% recall, and 100%, 77.8%, 100%, 90%, 95.7%, and 91.7% precision for SHSW, SHTW, walk, standing, BHSW, and BHTW actions, respectively.

Table 7. Classification matrix for Lena images.

	SHSW	*SHTW*	*Walk*	*Standing*	*BHSW*	*BHTW*	*Total Classified Images*	*Recall*	*Commission Error*	*Un-Classified Images*	*Total Images*	*Un-Classification*
SHSW	**24**	0	0	0	0	0	24	100%	0%	2	26	7.7%
SHTW	0	**21**	0	2	0	0	23	91.3%	8.7%	0	23	0%
Walk	0	0	**67**	2	0	0	69	97.1%	2.9%	1	70	1.4%
Standing	0	6	0	**36**	0	0	42	85.7%	14.3%	3	45	6.7%
BHSW	0	0	0	0	**22**	2	24	91.7%	8.3%	2	26	7.7%
BHTW	0	0	0	0	1	**22**	23	95.7%	4.3%	1	24	4.2%
Total Images	24	27	67	40	23	24	n = 205			9	m = 214	Overall Un-Classification(4.2%)
Precision	100%	77.8%	100%	90.0%	95.7%	91.7%	Overall Accuracy(93.7%)					
Omission Error	0.0%	22.2%	0%	10.0%	4.3%	8.3%						

Table 8 is the representation of the classification matrix for the actions performed by the actor "Lyova". There were 189 images for all the six actions from which 12 images remained unclassified, while for the remaining 177 images the overall accuracy of the proposed technique is 92.7%. The results achieved from the proposed technique are 95.2%, 85.7%, 95.7%, 89.6%, 100%, and 94.1% recall, and 100%, 80%, 100%, 86%, 100%, and 100% precision for for SHSW, SHTW, walk, standing, BHSW, and BHTW actions, respectively.

Table 8. Classification matrix for Lyova images.

	SHSW	*SHTW*	*Walk*	*Standing*	*BHSW*	*BHTW*	*Total Images Classified*	*Recall*	*Commission Error*	*Un-identified*	*Total Images*	*Un-Classification*
SHSW	**20**	1	0	0	0	0	21	95.2%	4.8%	3	24	12.5%
SHTW	0	**24**	0	4	0	0	28	85.7%	14.3%	2	30	6.7%
Walk	0	0	**44**	2	0	0	46	95.7%	4.3%	2	48	4.2%
Standing	0	5	0	**43**	0	0	48	89.6%	10.4%	4	52	7.7%
BHSW	0	0	0	0	**17**	0	17	100.0%	0.0%	1	18	5.6%
BHTW	0	0	0	1	0	**16**	17	94.1%	5.9%	0	17	0%
Total Images	20	30	44	50	17	16	n = 177			12	m = 189	Overall Un-Classification(6.3%)
Precision	100.0%	80.0%	100.0%	86.0%	100.0%	100.0%	Overall Accuracy(92.7%)					
Omision Error	0%	20.0%	0%	14.0%	0%	0%						

The classification matrix for the images of actions performed by "Moshe" is represented in Table 9. There were a total of 242 images, out of whom 11 images couldn't get classified through the proposed technique while the overall accuracy for the remaining 231 images is 87.4%. The precision and recall for SHSW, SHTW, walk, standing, BHSW, and BHTW are 100%, 61.9%, 100%, 75%, 100%, and 92.9%, and 89.3%, 86.7%, 97.3%, 71.7%, 90.5%, and 83.9% respectively.

Table 9. Classification matrix for Moshe images.

	SHSW	*SHTW*	*Walk*	*Standing*	*BHSW*	*BHTW*	*Total Images Classified*	*Recall*	*Commission Error*	*Un-Classified Images*	*Total Images*	*Un-Classification (%)*
SHSW	**25**	3	0	0	0	0	28	89.3%	10.7%	2	30	6.7%
SHTW	0	**26**	0	4	0	0	30	86.7%	13.3%	0	30	0.0%
Walk	0	0	**73**	2	0	0	75	97.3%	2.7%	1	76	1.3%
Standing	0	13	0	**33**	0	0	46	71.7%	28.3%	1	47	2.1%
BHSW	0	0	0	0	**19**	2	21	90.5%	9.5%	2	23	8.7%
BHTW	0	0	0	5	0	**26**	31	83.9%	16.1%	5	36	13.9%
Total Images	25	42	73	44	19	28	n = 231			11	m = 242	Overall Un-Classification(4.5%)
Precision	100%	61.9%	100%	75.0%	100%	92.9%	Overall Accuracy(87.4%)					
Omission Error	0%	38.1%	0%	25.0%	0%	7.1%						

Table 10 is representing the classification matrix for the complete dataset. The same as the classification matrices for the individual actors, it is showing the statistical results obtained by implementing the proposed technique for different parameters, namely; precision, recall, overall accuracy, overall Un-Classification, commission error, and omission error. The dataset contained 1855 images from all the actors having all the six actions. The proposed technique is unable to classify 84 actions while the remaining 1771 are classified with an overall accuracy of 91.4%. The recall values are 95.7%, 90.3%, 90.5%, 88.9%, 94.1%, and 92.3%, and the precision values are 98.6%, 84%, 97.7%, 81.3%, 99.4%, and 94.2% for SHSW, SHTW, walk, standing, BHSW, and BHTW, respectively.

Table 10. Classification matrix for the complete dataset.

	SHSW	*SHTW*	*Walk*	*Standing*	*BHSW*	*BHTW*	*Total Images Classified*	*Recall*	*Commission Error*	*Un-Classified Images*	*Total Images*	*Un-Classification (%)*
SHSW	**223**	10	0	0	0	0	233	95.7%	4.3%	16	249	6.4%
SHTW	3	**252**	0	24	0	0	279	90.3%	9.7%	13	292	4.5%
Walk	0	0	**428**	45	0	0	473	90.5%	9.5%	07	480	1.5%
Standing	0	38	7	**361**	0	0	406	88.9%	11.1%	13	419	3.1%
BHSW	0	0	0	0	**175**	11	186	94.1%	5.9%	19	205	9.2%
BHTW	0	0	0	14	1	**179**	194	92.3%	7.7%	16	210	7.6%
Total Images	226	300	435	444	176	190	n = 1771			84	m = 1855	Overall Un-Classification(4.5%)
Precision	98.6%	84.0%	97.7%	81.3%	99.4%	94.2%	Overall Accuracy(91.4%)					
Omission Error	1.4%	16.0%	2.3%	18.7%	0.6%	5.8%						

Figure 6 is showing the performance of the proposed technique using precision, recall, and F-score metrics. The statistics are presented for SHSW, SHTW, Walk, Standing, BHSW, and BHTW. The proposed technique has been applied over the Weizmann Dataset for all the performance parameters. In case of precision, it achieved 0.98, 0.84, 0.98, 0.81, 0.99, and 0.94, in case of recall; 0.96, 0.90, 0.91, 0.89, 0.94, and 0.92 while in the case of F-score; 0.97, 0.87, 0.94, 0.85, 0.97, and 0.93 have been achieved for SHSW, SHTW, walk, standing, BHSW, and BHTW actions, respectively. The highest precision has been achieved both in the case of SHSW and BHSW while standing has the least precision value of 0.81. The highest recall value of 0.96 has been achieved for SHSW action while the least recall has been recorded

for "standing" action having a value of 0.89. Actions of SHSW and BHSW achieved the highest F-score sharing 0.97 value while the least F-score is for "standing" action having a value of 0.85.

Figure 6. Performance of the proposed technique.

4.4. Discussion

As has already been discussed, the proposed technique has been used to classify six actions i.e., SHSW, SHTW, walk, standing, BHSW, and BHTW. For this purpose, a "modified dataset" has been used where images with the above-mentioned actions have been used containing the actions of eight different actors. Three of the actors are male while the rest are female. All of them have different heights, postures of walking, standing, and hands waving. The clear picture of the statistical results obtained from the proposed technique has been presented in Table 10 containing the classification matrix for the complete dataset.

The dataset originally contains 249 images for SHSW action. A total of 223 images for SHSW are correctly classified, 10 are misclassified as SHTW, while 16 actions are not classified by the proposed technique. The classification rule for the right SHSW has been shown in Equation (13). Most of the time, the first portion of Equation (13) i.e., $TL(\theta) > 125$ AND $TR(\theta) > 75$, get satisfied when the result is Un-Classification, but the second part of the rule i.e., $D_{TLR} > armLength$ is the cause of misclassification as if there is a bend in the arm or the actor does not have its complete stretch. In this case, D_{TLR} evaluates to less than the armLength resulting in it being classified out of the SHSW class. These actions don't even fall into the SHTW class as rule (5); the class does not get satisfied as the example image shown in Figure 7a. These are the cases where $TL(\theta)$ are greater than 125, but D_{TLR} is less than armLength or $TL(\theta) \leq 125$ but $armLength > D_{TLR} > (\frac{2}{3}) * armLength$. In all of these cases, SHSW actions are un-classified. Ten of SHSW actions that are misclassified as SHTW are those where the hand of the actor is in such a position that its feature D_{TLR} becomes less than $(\frac{2}{3}) * armLength$ as shown in Figure 7b. As described above, these are the cases where the arm is in a bent position or it is a side wave, but the position of the hand is above the normal side hand wave position.

(a) (b)

Figure 7. Abnormal "SHSW" actions. (**a**) Misclassification example; (**b**) Un-Classification example.

SHTW is the second action in the sequence whose statistics are shown in Table 10. The total number of images for SHTW actions from all the actors is 292. Out of all the 292 images, 252 are correctly classified. The Un-classified SHTW actions are 13 while the remaining images are misclassified as SHSW (03), and standing (24). Three of the SHTW actions are classified as SHSW. As discussed earlier in the case of SHSW, these are the actions where the actors stretched their arm more than the normal of the SHTW position and elbow bent was missing. As the SHTW rule for the right hand is defined in (15) and these cases fail to satisfy $D_{TLR} < (\frac{2}{3}), *armLength$ fulfills the condition of SHSW i.e., $D_{TLR} > armLength$. The other misclassification class is standing. In some of the Images of SHTW actions, the hand of the actors was just touching the head and even it was not much above the head failing to fulfill both the conditions of rule (15) and shown in Figure 8a i.e., $TL(\theta) > 97.5$ AND $TL(\theta) \leq 125)$ AND $TR(\theta) > 75$ and $D_{TLR} > hand\_Size$ AND $D_{TLR} < (\frac{2}{3}) * armLength$. As the top hand of the actor touches his/her head, the $TL(\theta)$ becomes less than 97.5 and D_{TLR} is not greater than hand_size so it matches the conditions of rule (1). In the SHTW Un-Classification cases as already discussed, these are the cases that neither fulfill rule (13) nor rule (15) as shown in Figure 8b.

(a) (b)

Figure 8. Abnormal "SHTW" actions. (**a**) Misclassification example; (**b**) Un-Classification example.

Classification statistics of "Walk" as action are presented in Table 10 showing cumulative results from all the actors. A total of 480 images having a person in walk posture have been used for validating the proposed technique. The number of classified images is 473 while 07 images remained un-classified. The images which are correctly classified are 428 while 45 actions are misclassified. All of the 45 actions are misclassified as Standing. Rule (12) is defined for classifying "Walk" action i.e., $D_{BLR}(\theta) > 15$ AND $SHR > 0.25$. The second part of the rule says that the step size to height ratio should be greater than 0.25. During the walk, the size of step changes and misclassification case is for those frames where the posture of step is such that it is getting towards 0. Rule (11) is defined for a Standing case whose second part says $SHR < 0.25$. Thus, as the step size of the person gets

shorter, therefore, the posture matches the standing posture and it may be taken as misclassification. The example images of those are shown in Figure 9a,b.

(a) (b)

Figure 9. Abnormal "Walk" actions. (**a**,**b**) Misclassification examples.

"Standing" is the next action whose classification statistics are shown in Table 10. A total of 419 images are used for testing purposes which are contributed by all of the eight actors. The number of classified images is 406 while 13 are un-classified. From the classified images, 361 are correctly classified while 45 are misclassified. Out of 45, 38 are misclassified as SHTW and seven as Walk. The classification rule for "Standing" action is given in (11) stating that three of the conditions need to be fulfilled for the prescribed class. The misclassification to SHTW is unfulfillment of portion of (11) i.e., $d_{TLR}(\theta) < 10$ AND $D_{TLR} \leq handSize$. These are the images in the dataset where the actor is standing in a bent position. The bent is more than the normal position and is thus closer to bending action than standing. This posture e.g., in Figure 10a resulted when D_{TLR} greater than handSize as the Top left point moves farther from the top right point. The second condition also gets failed i.e., $d_{TLR}(\theta) < 10$. As a result, rule (15) gets applied which classifies the action as SHTW. The second case where the images are misclassified as "Walk" doesn't fulfill the second half of rule (11) i.e., $SHR < 0.25$. Figure 10b shown below is the example of those misclassified action images. It is clear from Figure 10b that the position of feet of the actor is such that it gives the same "Walk" like step i.e., $SHR > 0.25$, which is a condition of rule (12) for "Walk" action.

(a) (b)

Figure 10. Abnormal "Standing" actions. (**a**,**b**) Misclassification examples.

The next action under discussion is "BHSW" for which a total of 205 test images are collected from eight different actors. The classified images are 186 while 19 of them got un-classified. From 186 classified actions, 175 are correctly classified and the remaining 11 are misclassified to "BHTW". The classification rule for BHSW is defined in (17). The classification rule for BHSW has two parts i.e., (i) Condition fulfilling top left and top right angles with respect to centroid of the human body, $TL(\theta) > 125$ AND $TR(theta) < 55$, (ii) The distance between top left and top right points, $D_{TLR} > 1.5 * armLength$. Figure 11a is an example of misclassification of "BHSW" as "BHTW" as both $TL(\theta)$

and $TR(theta)$ do not fulfill the required criteria and the conditions fall in rule (18) which is for "BHTW". The "BHSW" actions that are un-classified by the proposed technique neither fulfill rule (17) for "BHSW" nor rule (18) for "BHTW". Figure 11b,c are examples of un-classified actions. Figure 11b doesn't fulfill the first part of rule (17) so it can't be classified to "BHSW", although the first part of rule (18) is true i.e., $(TL(\theta) > 97.5$ AND $TL(\theta) \leq 125)$ AND $(TR(theta) < 82.5$ AND $TR(theta) \geq 55))$, but the second half of the rule i.e., $D_{TLR} \leq 1.5 * armLength$ is not fulfilled so resulting them as un-classified. The second Un-Classification example for "BHSW" is Figure 11c. This is the case where one of the arms is a side wave, while the position of the other is of the top wave. In these conditions, the first half of the rule (17) is partially true and rule (18) is not true at all so these actions are un-classified.

Figure 11. Abnormal "BHSW" actions. (**a**) Misclassification example; (**b**,**c**) Un-Classification examples.

The last action under the discussion is "BHTW". In Table 10, the classification statistical details for BHTW are presented as well. The contribution of eight actors resulted in 210 "BHTW" images. The classification rule for BHTW action is presented as (18). The number of images where the proposed technique did classification are 194, while the other 16 remained un-classified. Out of 194, 179 are correctly classified leaving 14 misclassified as "Standing" and one as "BHSW". The "BHTW" actions are misclassified as "Standing" as they are not able to fulfill the two halves of (18) i.e., (i) $(TL(\theta) > 97.5$ AND $TL(\theta) \leq 125)$ AND $(TR(theta) < 82.5$ AND $TR(theta) \geq 55))$ and (ii) $(D_{TLR} > forearm$ AND $D_{TLR} \leq 1.5 * armLength))$. Figure 12b is the example of the "BHTW" action images where left and right hands touch each other. When the top left and top right points are calculated, they do not fulfill the $D_{TLR} > forearm$ condition but do fulfill criteria from rule (11) i.e., $D_{TLR} \leq handSize$ for "Standing" action, so Figure 12b is classified as "Standing". Figure 12c is the image which is misclassified as "BHSW" as the part of rule (18) is not fulfilled i.e., $D_{TLR} \leq 1.5 * armLength$, but it agrees with all the parts of the rule defined in (17) resulting in being classified as "BHSW". Figure 12a is the example where neither the conditions of rules (18), (17), and (11) are fulfilled nor any other rule cover feature statistics, so images like Figure 12a remain un-classified.

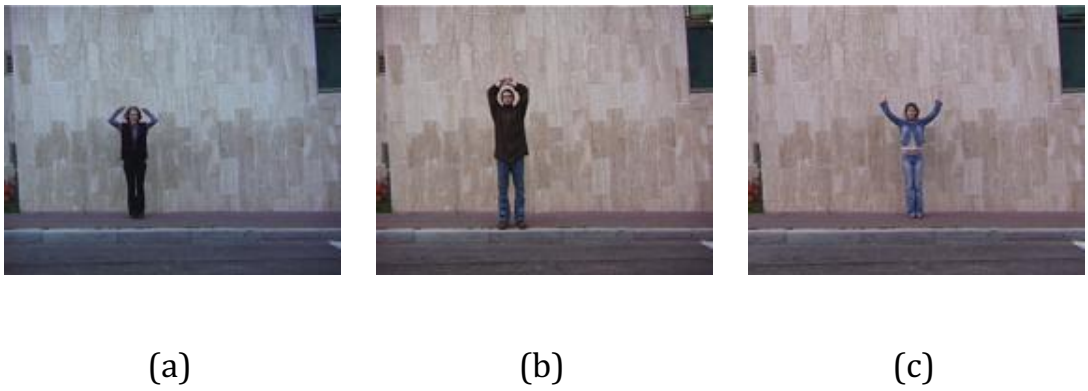


Figure 12. Abnormal "BHTW" Actions. (**a**) Un-Classification example; (**b**,**c**) Misclassification examples.

4.4.1. Comparative Analysis of the Proposed Technique with Existing Research

Figure 13 is showing comparative results of the proposed technique with the existing research. The comparison is based on the evaluation metric of overall accuracy. It may be observed that the accuracy of linear regression-based classification [23] over the Weizmann dataset is 61.7%, while it is 60.1% for subspace discriminant analysis [23]. The proposed technique achieved an overall accuracy of 91.4%, and it is highest among all of its counterparts. The results of multiclass SVMs [23] are 91.2% accurate while KNN based classification [23] was 86.1% correct. Dollár et al. [21] achieved 85.2% accuracy using sparse spatio-temporal features and unsupervised learning based classification through spatio-temporal words [46] attained an overall accuracy of 90.0%. Again, the spatial-temporal features were used by Klaser et al. [20], but, based on 3D gradients, it remained 84.3% accurate. It may also be observed that the Gaussian mixture model-based technique [18] could attain 91.11% accuracy and the same research using the nearest neighbor classifier [18] remained 87.78% accurate. Even the latest research [19] achieved an accuracy of 89.41%.

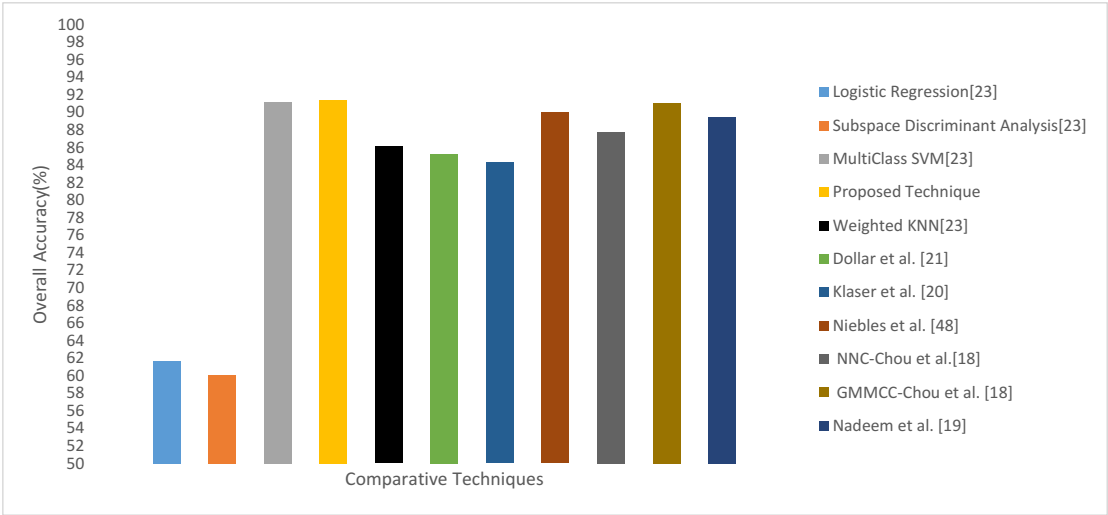


Figure 13. Accuracy based comparative results of the proposed technique with existing research.

5. Conclusions

A geometric featured based technique, where features are extracted in the context of human body science, is presented here. The proposed technique recognized six human actions, i.e., standing, walking, single-hand side wave, single-hand top wave, both hands side wave, and both hands top wave. All these actions are represented by using extreme points of the human body in each of the four

quadrants. The centroid of the human blob is also computed, allowing a relative calculation to be made about it. The dimensions of the human body (arm size, height, wingspan, hand size, and step-to-height ratio) are used in the classification rules and the results are presented in the form of classification matrices. BHSW having the highest precision of 99.4%, while "standing" has the least precision value of 81.3%. The highest recall is 95.7% for SHSW action while the least is for the "standing" action with 88.9%. Both SHSW and BHSW shared the highest F-score value of 97%, while the "standing" action has the least F-score value of 85%. In comparison to the existing research, the proposed technique remained at the top having 91.4% accuracy. In the future, the work may be extended for more complex situations where actions are completed through the participation of more than one human.

Author Contributions: Conceptualization, S.M.S.; Data curation, I.K.; Formal analysis, S.M.S., A.G., S.A.K.G. and S.S.; Funding acquisition, S.S. and N.N.; Investigation, S.M.S.; Methodology, S.M.S. and I.K.; Project administration, S.A.K.G.; Resources, A.G., S.A.K.G. and N.N.; Software, S.M.S. and I.K.; Validation, A.G. and S.A.K.G.; Visualization, I.K. and N.N.; Writing—original draft, S.M.S.; Writing—review & editing, A.G., S.S. and N.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We acknowledge the support of the German Research Foundation (DFG) and the Bauhaus-Universität Weimar within the Open-Access Publishing Programme.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
2. Zhu, Q.; Yeh, M.C.; Cheng, K.T.; Avidan, S. Fast human detection using a cascade of histograms of oriented gradients. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Washington, DC, USA, 17–22 June 2006; Volume 2, pp. 1491–1498.
3. Pang, Y.; Yuan, Y.; Li, X.; Pan, J. Efficient HOG human detection. *Signal Process.* **2011**, *91*, 773–781. [CrossRef]
4. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the 29th Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
5. Rebbah, S.; Nicol, F.; Puechmorel, S. The geometry of the generalized gamma manifold and an application to medical imaging. *Mathematics* **2019**, *7*, 674. [CrossRef]
6. Hervella, Á.S.; Rouco, J.; Novo, J.; Ortega, M. Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement. *Procedia Comput. Sci.* **2018**, *126*, 97–104. [CrossRef]
7. Evangelidis, G.; Singh, G.; Horaud, R. Skeletal quads: Human action recognition using joint quadruples. In Proceedings of the International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 4513–4518.
8. Zhang, J.; Lin, H.; Nie, W.; Chaisorn, L.; Wong, Y.; Kankanhalli, M.S. Human action recognition bases on local action attributes. *J. Electr. Eng. Technol.* **2015**, *10*, 1264–1274. [CrossRef]
9. Marín-Jiménez, M.J.; de la Blanca, N.P.; Mendoza, M.A. Human action recognition from simple feature pooling. *Pattern Anal. Appl.* **2014**, *17*, 17–36. [CrossRef]
10. Al-Ali, S.; Milanova, M.; Al-Rizzo, H.; Fox, V.L. Human Action Recognition: Contour-Based and Silhouette-Based Approaches. In *Computer Vision in Control Systems-2*; Springer: Berlin, Germany, 2015; pp. 11–47.
11. Veenendaal, A.; Daly, E.; Jones, E.; Gang, Z.; Vartak, S.; Patwardhan, R.S. Decision Rule Driven Human Activity Recognition. *Comput. Sci. Emerg. Res. J.* **2015**, *3*, 1–7.
12. Wu, B.; Yuan, C.; Hu, W. Human action recognition based on context-dependent graph kernels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 2609–2616.
13. Veenendaal, A.; Jones, E.; Gang, Z.; Daly, E.; Vartak, S.; Patwardhan, R. Dynamic Probabilistic Network Based Human Action Recognition. *arXiv* **2016**, arXiv:1610.06395.

14. Abdulmunem, A.; Lai, Y.K.; Sun, X. Saliency guided local and global descriptors for effective action recognition. *Comput. Vis. Media* **2016**, *2*, 97–106. [CrossRef]
15. Liang, F.; Zhang, Z.; Li, X.; Tong, Z. Lower Limb Action Recognition with Motion Data of a Human Joint. *Arab. J. Sci. Eng.* **2016**, *41*, 5111–5121. [CrossRef]
16. Luvizon, D.C.; Tabia, H.; Picard, D. Learning features combination for human action recognition from skeleton sequences. *Pattern Recognit. Lett.* **2017**, *99*, 13–20. [CrossRef]
17. Saggese, A.; Strisciuglio, N.; Vento, M.; Petkov, N. Learning skeleton representations for human action recognition. *Pattern Recognit. Lett.* **2019**, *118*, 23–31. [CrossRef]
18. Chou, K.P.; Prasad, M.; Wu, D.; Sharma, N.; Li, D.L.; Lin, Y.F.; Blumenstein, M.; Lin, W.C.; Lin, C.T. Robust feature-based automated multi-view human action recognition system. *IEEE Access* **2018**, *6*, 15283–15296. [CrossRef]
19. Nadeem, A.; Jalal, A.; Kim, K. Human Actions Tracking and Recognition Based on Body Parts Detection via Artificial Neural Network. In Proceedings of the 2020 3rd International Conference on Advancements in Computational Sciences (ICACS), Berlin, Germany, 26–28 September 2020; pp. 1–6.
20. Klaser, A.; Marszałek, M.; Schmid, C. A Spatio-Temporal Descriptor Based on 3d-Gradients. In Proceedings of the British Machine Vision Conference(BMVP), Leeds, UK, 1–4 September 2008.
21. Dollár, P.; Rabaud, V.; Cottrell, G.; Belongie, S. *Behavior Recognition via Sparse Spatio-Temporal Features*; VS-PETS: Beijing, China, 2005.
22. Shah, S.M.S.; Malik, T.A.; Khatoon, R.; Hassan, S.S.; Shah, F.A. Human Behavior Classification Using Geometrical Features of Skeleton and Support Vector Machines. *CMC-Comput. Mater. Contin.* **2019**, *61*, 535–553. [CrossRef]
23. Sharif, M.; Khan, M.A.; Akram, T.; Javed, M.Y.; Saba, T.; Rehman, A. A framework of human detection and action recognition based on uniform segmentation and combination of Euclidean distance and joint entropy-based features selection. *EURASIP J. Image Video Process.* **2017**, *2017*, 89. [CrossRef]
24. Yoon, S.M.; Kuijper, A. Human action recognition based on skeleton splitting. *Expert Syst. Appl.* **2013**, *40*, 6848–6855. [CrossRef]
25. Ran, X.; Xue, L.; Zhang, Y.; Liu, Z.; Sang, X.; He, J. Rock Classification from Field Image Patches Analyzed Using a Deep Convolutional Neural Network. *Mathematics* **2019**, *7*, 755. [CrossRef]
26. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 20–36.
27. Zhu, W.; Lan, C.; Xing, J.; Zeng, W.; Li, Y.; Shen, L.; Xie, X. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. *arXiv* **2016**, arXiv:1603.07772.
28. Chéron, G.; Laptev, I.; Schmid, C. P-CNN: Pose-based CNN features for action recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 3218–3226.
29. Pan, X.; Guo, W.; Guo, X.; Li, W.; Xu, J.; Wu, J. Deep Temporal–Spatial Aggregation for Video-Based Facial Expression Recognition. *Symmetry* **2019**, *11*, 52. [CrossRef]
30. Kataoka, H.; Miyashita, Y.; Hayashi, M.; Iwata, K.; Satoh, Y. Recognition of transitional action for short-term action prediction using discriminative temporal cnn feature. In Proceedings of the British Machine Vision Conference (BMVC), York, UK, 19–22 September 2016.
31. Chen, C.; Jafari, R.; Kehtarnavaz, N. Action recognition from depth sequences using depth motion maps-based local binary patterns. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2015; pp. 1092–1099.
32. Chen, C.; Jafari, R.; Kehtarnavaz, N. Fusion of depth, skeleton, and inertial data for human action recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2712–2716.
33. Li, M.; Leung, H.; Shum, H.P. Human action recognition via skeletal and depth based feature fusion. In Proceedings of the 9th International Conference on Motion in Games, Burlingame, CA, USA, 10–12 October 2016; pp. 123–132.
34. Jiang, M.; Kong, J.; Bebis, G.; Huo, H. Informative joints based human action recognition using skeleton contexts. *Signal Process. Image Commun.* **2015**, *33*, 29–40. [CrossRef]
35. Chaaraoui, A.A.; Padilla-López, J.R.; Climent-Pérez, P.; Flórez-Revuelta, F. Evolutionary joint selection to improve human action recognition with RGB-D devices. *Expert Syst. Appl.* **2014**, *41*, 786–794. [CrossRef]

36. Kim, H.; Lee, S.; Kim, Y.; Lee, S.; Lee, D.; Ju, J.; Myung, H. Weighted joint-based human behavior recognition algorithm using only depth information for low-cost intelligent video-surveillance system. *Expert Syst. Appl.* **2016**, *45*, 131–141. [CrossRef]
37. Chaaraoui, A.A.; Climent-Pérez, P.; Flórez-Revuelta, F. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognit. Lett.* **2013**, *34*, 1799–1807. [CrossRef]
38. Guo, G.; Lai, A. A survey on still image based human action recognition. *Pattern Recognit.* **2014**, *47*, 3343–3361. [CrossRef]
39. Zhao, Z.; Huimin, M.; Xiaozhi, C. Generalized symmetric pair model for action classification in still images. *Pattern Recognit.* **2017**, *64*, 347–360. [CrossRef]
40. Sharma, G.; Jurie, F.; Schmid, C. Expanded parts model for human attribute and action recognition in still images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 652–659.
41. Vishwakarma, D.; Kapoor, R. Hybrid classifier based human activity recognition using the silhouette and cells. *Expert Syst. Appl.* **2015**, *42*, 6957–6965. [CrossRef]
42. Plagenhoef, S.; Evans, F.G.; Abdelnour, T. Anatomical data for analyzing human motion. *Res. Q. Exerc. Sport* **1983**, *54*, 169–178. [CrossRef]
43. Elert, G. As Size of a Human: Body Proportions. The Physics Factbook. 2006. Available online: http://hypertextbook.com/facts/2006/bodyproportions.shtml (accessed on 25 July 2019).
44. Elert, G. The Physics Factbook; 2000. Available online: http://hypertextbook.com/facts/index-topics.shtml (accessed on 25 July 2019).
45. Blank, M.; Gorelick, L.; Shechtman, E.; Irani, M.; Basri, R. Actions as space-time shapes. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), Piscataway, NJ, USA, 17–21 October 2005; Volume 2, pp. 1395–1402.
46. Niebles, J.C.; Wang, H.; Fei-Fei, L. Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vis.* **2008**, *79*, 299–318. [CrossRef]