symmetry

MDPI

Article

# Accurate Physical Activity Recognition using Multidimensional Features and Markov Model for Smart Health Fitness

**Amir Nadeem [1], Ahmad Jalal [1] and Kibum Kim [2,*]**

[1] Department of Computer Science, Air University, Islamabad 44000, Pakistan; 171269@students.au.edu.pk (A.N.); ahmadjalal@mail.au.edu.pk (A.J.)
[2] Department of Human-Computer Interaction, Hanyang University, Ansan 15588, Korea
* Correspondence: kibum@hanyang.ac.kr

Received: 7 August 2020; Accepted: 22 October 2020; Published: 24 October 2020


**Abstract:** Recent developments in sensor technologies enable physical activity recognition (PAR) as an essential tool for smart health monitoring and for fitness exercises. For efficient PAR, model representation and training are significant factors contributing to the ultimate success of recognition systems because model representation and accurate detection of body parts and physical activities cannot be distinguished if the system is not well trained. This paper provides a unified framework that explores multidimensional features with the help of a fusion of body part models and quadratic discriminant analysis which uses these features for markerless human pose estimation. Multilevel features are extracted as displacement parameters to work as spatiotemporal properties. These properties represent the respective positions of the body parts with respect to time. Finally, these features are processed by a maximum entropy Markov model as a recognition engine based on transition and emission probability values. Experimental results demonstrate that the proposed model produces more accurate results compared to the state-of-the-art methods for both body part detection and for physical activity recognition. The accuracy of the proposed method for body part detection is 90.91% on a University of Central Florida's (UCF) sports action dataset and, for activity recognition on a UCF YouTube action dataset and an IM-DailyRGBEvents dataset, accuracy is 89.09% and 88.26% respectively.

**Keywords:** body parts detection; Markov model; physical activity recognition; spatiotemporal features

## 1. Introduction

Assistive technologies for human locomotion tracking provide independent mobility, social participation and health benefits [1]. These benefits have emerged as a major research gain in worldly application domains such as violence detection, home automation systems, customer surveillance, virtual reality and physical fitness [2,3]. However, the tracking and recognition of people's physical activities remain problematic due to the human body's articulated nature, degrees of freedom between joints, partial occlusion and varying scales normalization [4]. Several modules such as rigid body configuration, body-part landmarks, homograph estimation, and optimal feature descriptors are introduced to minimize these difficulties.

Although, a lot of efforts have been put in by researchers in physical activity recognition (PAR), some challenges are still unresolved as described below:

1. Shape and height variations: human size and shape appear smaller when individuals are further away from the camera; when they are closer they appear larger. In addition, human bodies vary a lot in shape and size.

2. Feature selection approach: there is a huge number of feature selection approaches. To choose an appropriate approach for feature selection for the PAR is a critical issue.
3. Occlusion: a human body or any part of a particular body may be hidden due to occlusion.
4. Hardware problems: many approaches for PAR use expensive hardware in their research making it difficult to incorporate these systems in real life.
5. Illumination variations: the same image can look entirely different in different lighting situations.

In recent years, the interest of researchers in PAR has increased due to its numerous applications. Major application areas of PAR include video surveillance, virtual reality gaming, human–object interaction, e-learning, healthcare systems and human behavior analysis. In security surveillance systems, if a person walks normally under video surveillance and suddenly behaves suspiciously, there is a chance that abnormal events such as threats, fighting, domestic violence or agitation have occurred. Such abnormal activities are automatically detected by PAR to initiate the securing of the areas under surveillance. Similarly, health-exercise systems use automatic action-recognition technologies that can guide patients to exercise properly and assist them in their daily routines. In addition, PAR can make sports and games more attractive and entertaining due to the prediction of future players' actions and the expected scores of each team.

In this paper, we present a novel technique to estimate pose and to identify the specific physical activities based on a 12-point skeletal model and multidimensional features. These features are further reduced, optimized and classified by quadratic discriminant analysis (QDA), along with a maximum entropy Markov model (MEMM). The proposed method was tested on University of Central Florida's UCF sports actions, UCF YouTube actions and IM-DailyRGBEvent datasets (a collection of video sequences containing different common human actions) for body-part detection along with PAR. The test results achieved remarkable performance scores.

The rest of the paper is organized as follows. Section 2 consists of related work in the field of PAR. In Section 3, the framework is outlined, including system design, preprocessing stage, feature generation, and activity training/recognition. In Section 4, experimental results for body part detection and PAR are described. Finally, Section 5 describes the conclusion of our proposed work and future directions.

## 2. Related Work

Two types of sensors are commonly used in PAR. The first type is inertial sensors. Some gadgets (e.g., a smart watch) are worn by users at different locations on their body. Here, various accelerometers are embedded in the smart watch to measure acceleration forces. As activities are carried out, different acceleration forces are stored as data, preprocessing steps are performed and all activities are categorized. The second type of PAR sensors is vision sensors that recognize activities based on captured images. For PAR-based vision sensors, image sequences or video are captured by still/movable cameras and fed into the detector engine. Research has been done in PAR using both types of sensors and it is discussed in the following subsections.

### *2.1. Body Worn Sensors for PAR Systems*

In body worn inertial sensor based work, Trung et al. [5] recognized the similar types of actions that are usually difficult to classify. They used inter-class relationships to improve the overall performance of the method. In [6], Trung et al., used inertial sensors to recognize actions and, instead of using interclass relationships, they used a scale space method to segment the action signals properly. They also tackled the problem of inconsistency in sensor orientation by adjusting the tilt of the sensors. In both [5] and [6], despite using different methodologies, accuracy was low, so in [7], Hawang et al. suggested a new method for physical activity recognition by fusing an inertial sensor with a vision sensor to overcome the problem of unreliability of the inertial sensors. They claimed that merging these two types of sensors could help overcome deficiencies in both types of systems. In [8], Irvin and Angelica,

by contrast with [7], preferred to attach more inertial sensors to the user's body for PAR over the fusion of vision and inertial sensors. They estimated the angle of lower and upper limbs to extract features related to movement of the limbs. Dawar et al. [9] supported the fusion of sensors as in [7]. Additionally, they used convolutional neural networks to detect the data that was captured by vision sensors as well as another network "short term memory" with accelerometer data.

### 2.2. Vision Sensors for PAR Systems

In vision sensors, Fang et al. [10] developed a classical statistical sampling scheme along with deep learning representation of individual silhouettes to identify complex human actions. In [11], Silambarasi et al. developed a 3D spatio-temporal plane which locates human movements from different views via motion history tracing. They extracted histograms of oriented gradients and directional gradients and magnitudes to recognize physical activities. In [12], Shehzed et al. presented a new multiperson tracking system that included body-part labeling, Kalman filter and Gaussian mapping for crowd counting and action detection. In [13], Han et al. proposed a global spatial attention (GSA) model that explored different skeletal joints and adopted an accumulative learning curve to distinguish and recognize various action types. However, these articles [10–13] still have major issues such as uncontrolled lighting, dynamic postures, rotational views and motion ambiguities, which result in low performance. Therefore, to overcome these limitations, we developed a novel methodology for PAR.

## 3. Proposed System Methodology

We utilize video sensors to capture raw data during preprocessing; human silhouettes are extracted using two significant models including saliency based segmentation and skin tone detection; and then, these silhouettes are used to extract multilevel features including displacement parameter values. Finally, features are quantified using quadratic discriminant analysis (QDA) to get the best matching unit and to find maximum entropy of each activity class via a Markov model. Figure 1, depicts the proposed framework of our PAR system.

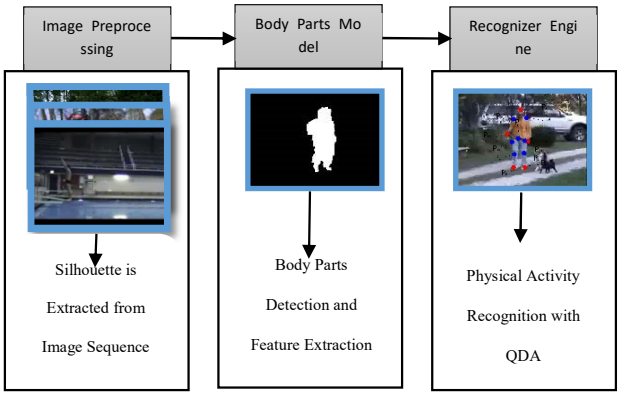


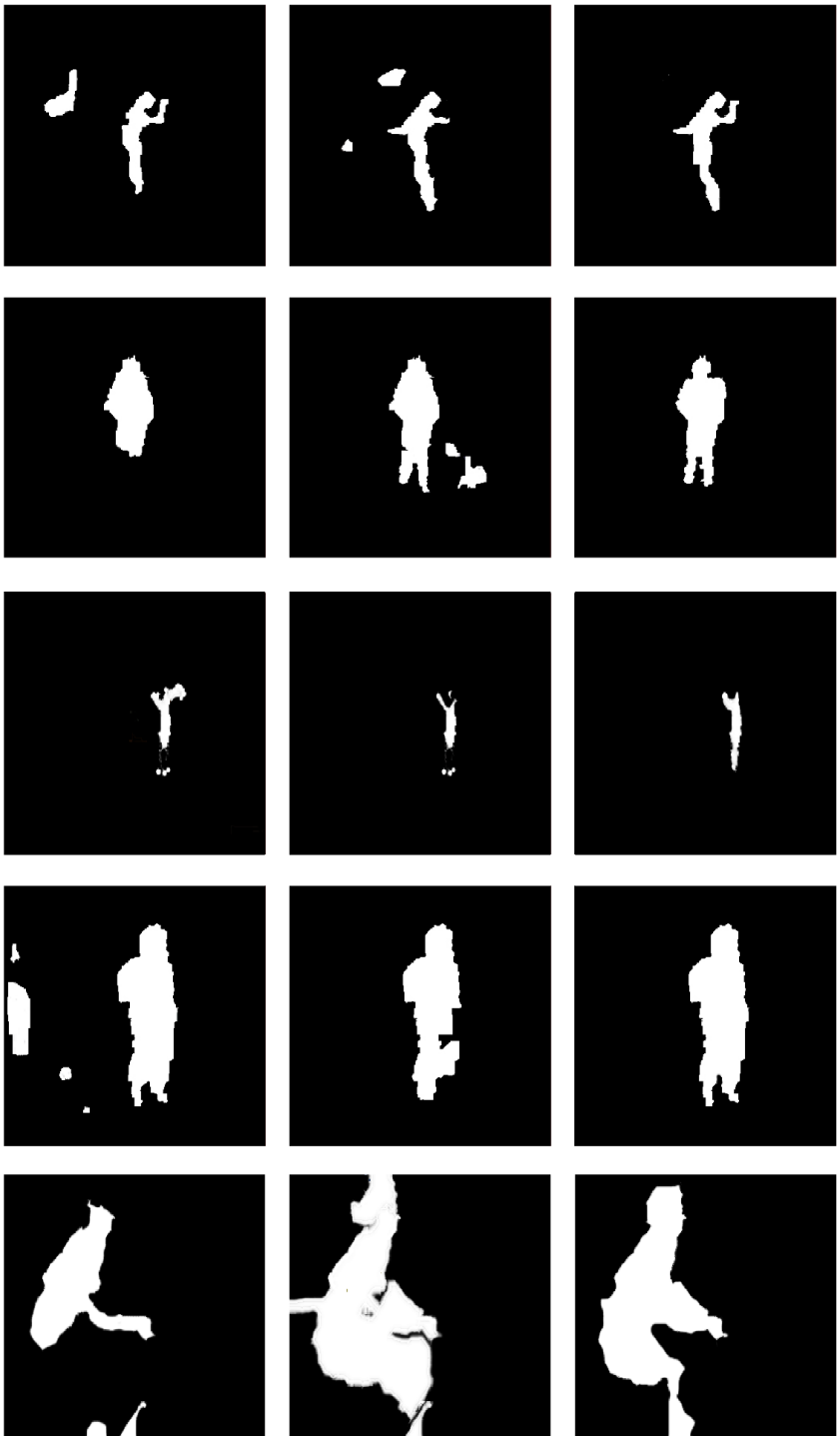

**Figure 1.** Overview of the proposed system architecture.

### 3.1. Preprocessing Stage

During vision-based image preprocessing, we applied two significant methods to extract reasonable human silhouettes. First, we extracted the silhouette using a salient region detection technique and then a separate silhouette was extracted using a skin tone segmentation technique. After the extraction of these two silhouettes, results were merged to get robust and accurate silhouettes from the given image. Saliency based segmentation [14] was used to distinguish an object (i.e., silhouette) by saliency values which were calculated from its surroundings. Saliency SR for pixel $(i,\ j)$ was computed as;

$$SR(x,y) = \sum_{(p,q)\in N} d[R(x,y), Q(p,q)] \quad (1)$$

$N$ is defined as an area near the saliency pixel at location $(x, y)$ and d is defined as the position difference between pixel vectors $R$ and $Q$. After determining the saliency values for all the given regions of the image, a standard threshold saliency value was used to differentiate foreground from background. Figure 2a shows the results of the saliency method.



(**a**) (**b**) (**c**)

**Figure 2.** Silhouette extraction methods. (**a**) Saliency base segmentation, (**b**) skin tone detection and (**c**) combined effect of both (**a**, **b**) methods.

The silhouette that was extracted with the skin tone approach [15] was used to improve the results of the saliency value method. In the skin tone approach, certain color ranges having specified decision borders were perceived. On the other hand, RGB and `hue, saturation, value` (HSV) threshold values were used to separate the skin region from the nonskin region. Equation (2) represents the RGB threshold, while Equation (3) was used to represent the HSV threshold for the skin tone segmentation.

$$r\_r = [0.36, 0.46], g\_r = [0.28, 0.35] \tag{2}$$

$$H_r = [0, 50], S_r = [0.20, 0.68] \; and \; V_r = [0.35, \; 1.0] \tag{3}$$

Here, $r\_r$ and $g\_r$ represent the range of red and green channels of RGB, respectively, whereas $H_r$, $S_r$ and $V_r$ represent the ranges of the HSV color model. After distinguishing the skin regions from nonskin regions, skin regions were elaborated with the help of the geometrical characteristics of the human body. In order to grow these regions, it was assumed that the skin was often visible only on the face, arms and lower legs. If two regions were detected vertically with the skin tone segmentation method, it was most likely that one of these regions was the face and the other was the lower legs. In that case, the system connected these regions. Thus, if further skin regions were found on the left/right side of the linked region, they may have been either hands or arms (See Figure 2b).

After extracting the silhouettes from each method, both methods were merged to get a more accurate silhouette. To combine both extraction methods, Algorithm 1 was formulated as:

**Algorithm 1:** Extraction of human silhouette

**Input:** Y: Result of saliency method, result of skin tone method
**Output**: Merged silhouette
/* Calculating position of human in image*/
/* β is denoting non zero region*/
/* μ is denoting merged silhouette*/
/* Ω is denoting human shape*/

**Step 1:**
**Repeat**
  **For** i=1 to N **do**
    **For** i=1 to N **do**
      $search(\beta)$
    **End**
  **End**
  **If** $\beta 1 > \beta$
    $\beta = \beta 1$
  **End**
**Until** biggest regions in both inputs are searched.

**Step 2:**
/* Compare β of both images*/
**For** all pixel in β of both images
  **If** $\beta_{pixel\ of\ image\ 1} = \beta_{pixel\ of\ image\ 2}$
    $\mu_{pixel\ of\ image\ 3} = \beta_{pixel\ of\ image\ 1}$
  **End**
  **If** β is unequal for both images
    **If** pixel is matching Ω
      $\mu_{pixel} = \beta_{pixel}$
    **End**
  **End**
**End**

### 3.2. Pose Estimation: Body Parts Detection

During pose estimation, the initial pose was considered as a T-shape with the arms extended straight out from the neck for human body configuration. Initially, five parts of the body were detected as basic parts [16]: the hands, the head, and the feet. An inflection-based method was incorporated in the system; it used the 2D kappa mechanism which was closely associated with object silhouette refining [17]. The Kappa function is defined as;

$$k(\gamma) = \frac{\dot{x}(\gamma)\ddot{y}(\gamma) - \dot{y}(\gamma)\ddot{x}(\gamma)}{\left(\dot{x}(\gamma)^2 + \dot{y}(\gamma)^2\right)^{3/2}} \tag{4}$$

After the silhouette was well refined with the Kappa function, the gap between the highest pixel of the silhouette and the lowest pixel of the silhouette was measured. The head diameter was standardized as 1/4.5th times the height to estimate the individual's height and width. In addition, taking into consideration pixelwise digging, the head diameter was calculated by measuring the altitude of the silhouette. To detect the head position, the following formula was used;

$$P_H^f \leftarrow P_H^{f-1} + \Delta P_H^{f-1} \tag{5}$$

where, $P_H^f$ is head position at any given frame $f$. The position of the limb was needed to estimate the locations of the hips and the feet (See Figure 3). The following equation was used to determine the limb position;

$$P_i^f = P_{i-1}^f + \left(r_{i-1}^{f-1} \ldots .. r_0^{f-1}\right).\left(P_i^{f-1} - P_{i-1}^{f-1}\right) \tag{6}$$

where, $P_i^f$ is the limb position in frame $i$. The position of the hands and feet were determined via lower and upper limb positions and the geometrical feature of the silhouette [18].

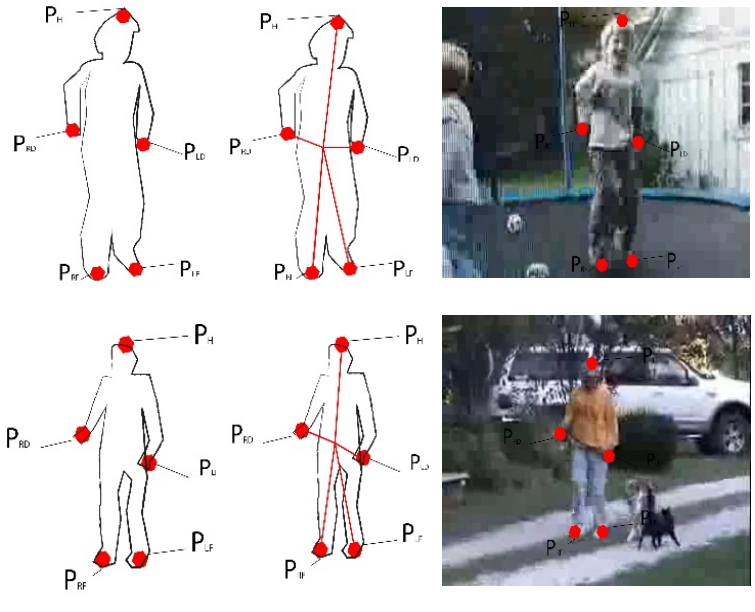



**Figure 3.** Basic body parts detection. (**a**) Five points are detected, (**b**) a star shape is extracted, and (**c**) shows body the parts in different poses.

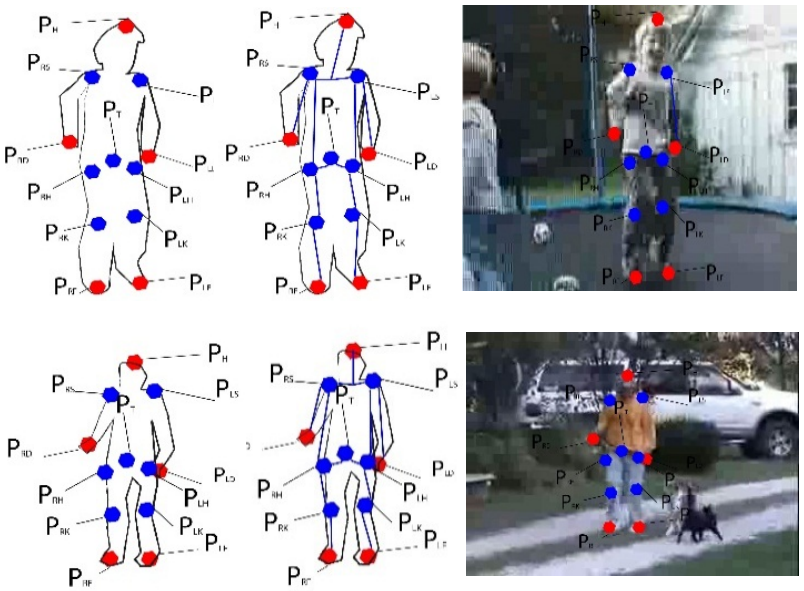
The torso point was at the center of the upper head point and between the feet [19]. The torso position was adjusted with the help of the following equation;

$$P_T^f \leftarrow P_T^{f-1} + \Delta P_T^{f-1} \tag{7}$$

Equation (8) was used to identify the location of the knees. It was usually at the center point between the feet and the hip joints.

$$P_K^f = \left(P_F^f - P_{Hip}^f\right)/2 \tag{8}$$

Thus we explored twelve body parts, being five basic body parts and seven body sub parts. As these images were in sequence we could track these parts and get optimal positions for each body part. Figure 4 gives a few examples of the detection of the 12 body parts.



(**a**) (**b**) (**c**)

**Figure 4.** Basic body parts detection. (**a**) Red points are the five basic body parts while the blue points highlight the seven body subparts, (**b**) skeleton structure with 12 points, and (**c**) 12 body parts at different poses.

### *3.3. Multidimensinal Features Generation*

Once the twelve body parts were detected from a human posture, we applied multilevel features. This included six dimensional torso features, eight dimensional first-degree features and eight dimensional second degree features. Algorithm 2 explains the overall definition of the multidimensional features.

### *3.4. Features Discrimination*

QDA [20] was used to evaluate which feature values can distinguish between all activity classes in labeled datasets. Each class was dispersed normally [21], and therefore a quantification function for quadratic discriminant analysis was applied as

$$D_i^2(x) = \left(x - \overline{x_j}\right)' S_j^{-1}\left(x - \overline{x_j}\right) \tag{9}$$

where, $S_j$ is covariance matrix and $j$ is $1, 2, \cdots k$. To distinguish between features, we examined if $D_i^2(x)$ is the smallest for any class x. Figure 5, shows a 3D plot having clear discrimination of 11 different classes of the UCF YouTube action dataset.

**Algorithm 2:** Feature Generation

**Input:** Y: Position of 12 body parts,
**Output**: Generated Features
/* η inclination angle*/
/* β is azimuth angle*/
/* R is radius*/
**6 Torso Features:**
/* Distance between head and torso, azimuth angle and inclination angle is used*/
$R = |P_T, P_H|$
$B = angle(azimuth\ axis\ , (P_T, P_H))$
$\eta = angle(zenith\ axis, (P_T, P_H))$
**8 First Degree Features:**
/* Distance from right hand to each body part, azimuth angle and inclination angle is used*/

**Repeat** for all body parts i
$R = |P_{RD}, P_i|$
$B = angle(azimuth\ axis\ , (P_{RD}, P_i))$
$\eta = angle(zenith\ axis, (P_{RD}, P_i))$
**END**
**8 Second Degree Features:**
/* Distance between left hand and each, azimuth angle and inclination angle is used*/
**Repeat** for all body parts i
$R = |P_{LD}, P_i|$
$B = angle(azimuth\ axis\ , (P_{LD}, P_i))$
$\eta = angle(zenith\ axis, (P_{LD}, P_i))$
**END**

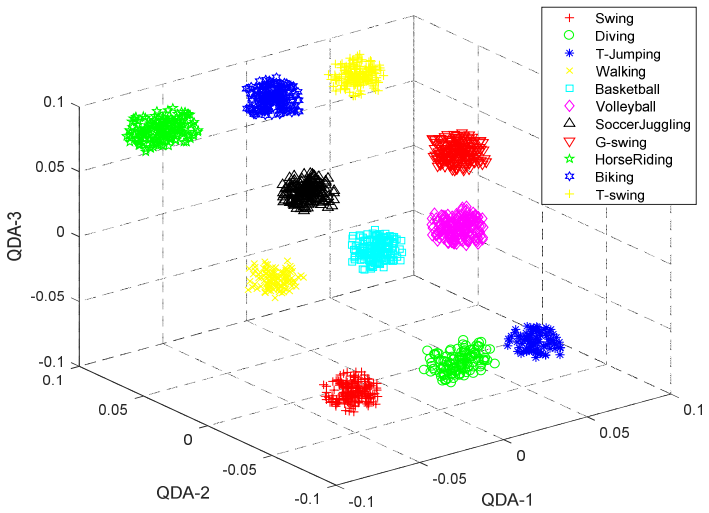



**Figure 5.** Discrimination of quadratic discriminant analysis (QDA) over all classes of the UCF YouTube action dataset.

Figure 6 shows the 3D plot with clear discrimination between 15 different classes over the IM-DailyRGBEvents dataset.

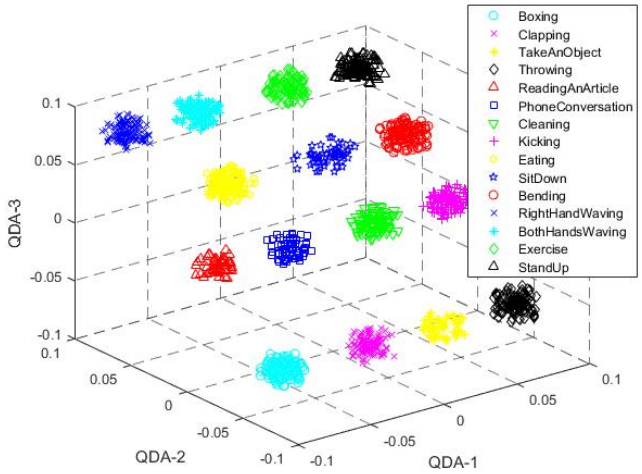



**Figure 6.** Discrimination of QDA over all classes of the IM-DailyRGBEvents dataset.

### 3.5. Recognition Engine: MEMM

For activity classification, conditional probability of the observation sequence was used to estimate the state sequence [22] via MEMM. According to the MEMM model [23], the activity classes were declared as the state $P\left(S_i \middle| S_{i-1},\ O_i\right)$ with entropy adjustments, formulated as:

$$P(S_1, \ldots, S_n | O_1, \ldots, O_n) = \mathrm{e}^{\sum_{k=1}^{x} \delta_k \beta_k} \tag{10}$$

where, $\delta_k$ is the feature value and $\beta_k$ is the adjustable weight $K$ for the given observation in the sequence. The conditional entropy of a distribution $P\left(S \middle| O\right)$ is estimated by maximum entropy theory. It was inferred by the log-linear model as:

$$P(S|O) = \frac{1}{Z\left(O, S'\right)} \exp\left(\sum\nolimits_m \lambda_m f_m(O, S)\right) \tag{11}$$

where $Z\left(O,\ S'\right)$ is a normalized factor and $\lambda_m$ is the multiplier parameter with multi-level features. Figure 7, shows how probability is estimated during MEMM classification over the different activities of *walking*, *swinging* and *T-jumping* in the YouTube action dataset.

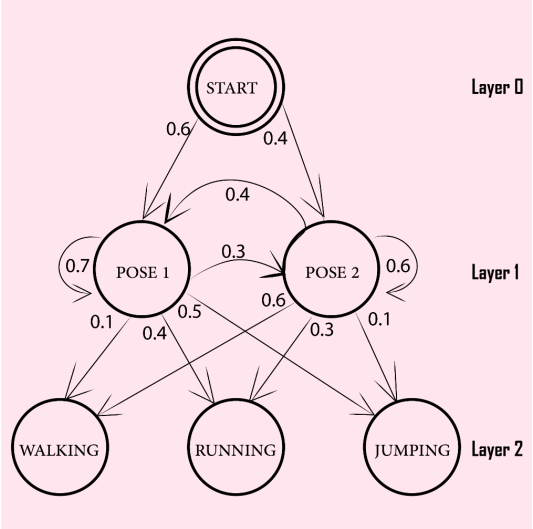



**Figure 7.** Maximum entropy Markov model (MEMM) classification of different classes using the UCF YouTube action dataset.

## 4. Experimental Results

In this section, firstly, we explain three different benchmark-challenging datasets. Four types of experimental results are represented after the explanation of three datasets. In the first experiment, we explored body part detection accuracies with respect to ground truth. In the second experiment, action, recognition accuracies are represented. In the third experiment, we compared the proposed technique with well-known machine learning algorithms. Finally, in the four experiment, we compared body part detection accuracies as well as action recognition accuracies with other statistical well-known state-of-the-art methods.

### *4.1. Datasets Description*

In the UCF sports actions dataset [24], a set of action classes was gathered from different games usually shown on TV stations like the BBC and the ESPN. The actions included *diving*, *golf swing*, *kicking*, *lifting*, *riding horse*, *running*, *skate boarding*, *swing-bench*, *swing-side* and *walking* of 720 × 480 resolution. The dataset is available as videos having more than a hundred sequences. Figure 8 shows some samples of the UCF sports actions dataset.

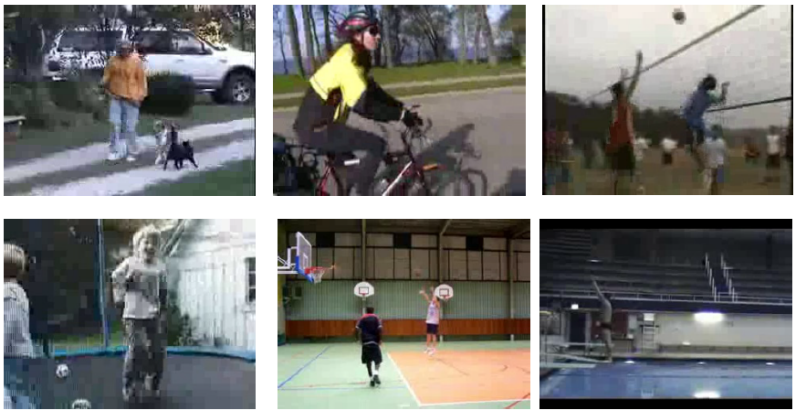


**Figure 8.** Some examples of UCF sports actions dataset.

In the UCF YouTube action [25] dataset, there were 11 different classes of action such as *swing*, *diving*, *T-jumping*, *walking*, *basketball*, *volleyball*, *soccer juggling*, *G-swing*, horse-*riding*, *biking*, and *T-swing*. The clips were combined into 25 groups per category, containing a minimum of four actions per clip. Videos of the same category shared common characteristics such as the same performer, common context and specific point of view. In Figure 9, there are some samples from the UCF YouTube action dataset.

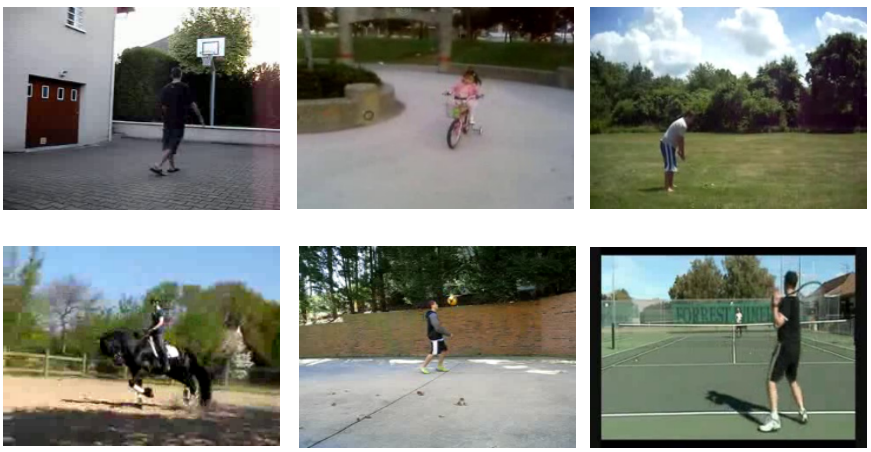


**Figure 9.** Some examples of UCF YouTube action dataset.

In our self-annotated IM-DailyRGBEvents dataset [26], there were 15 classes of actions performed by 15 subjects (i.e., 13 males, 2 females). There were more than seventy video sequences for each action. Figure 10, shows some images from the IM-DailyRGBEvents dataset.

**Figure 10.** A few samples of the IM-DailyRGBEvents dataset.

### *4.2. Experimentation I: Body Parts Detection Accuracies*

To calculate the effectiveness and accuracy of body part detection, the distance from the ground truth (GT) was calculated with the help of the following equation.

$$D = \sqrt{\sum_{k=1}^{K} \left( \frac{I_k}{S_k} - \frac{J_k}{S_k} \right)^2} \tag{12}$$

Here, $J$ is the GT and $I$ is the location of the detected body part. The threshold of 15 is set to identify accuracy between the detected data and the GT data. With the help of the following equation (13), the percentage of the detected parts that lies within the threshold range of labeled dataset was detected.

$$Detection\ Accuracy = \frac{100}{k} \left[ \sum_{k=1}^{K} \begin{cases} 1\ if\ D \leq 15 \\ 0\ if\ D > 15 \end{cases} \right] \tag{13}$$

In Table 1, column 2 is the distance from the ground truth and column 3 shows body part detection accuracy over the UCF sports action dataset.

**Table 1.** Accuracy of body parts detection.

| Body Parts | Distance from Ground Truth | Detection Accuracy (%) |
|---|---|---|
| **Upper head** | 11.3 | 98 |
| **Left Shoulder** | 9.9 | 95 |
| **Right Shoulder** | 13.6 | 90 |
| **Left hand** | 13.2 | 93 |
| **Right Hand** | 10.3 | 90 |
| **Left Hip** | 8.8 | 83 |
| **Right Hip** | 11.7 | 87 |
| **Left Knee** | 14.3 | 85 |
| **Right Knee** | 10.3 | 85 |
| **Left foot** | 9.8 | 97 |
| **Right foot** | 10.9 | 96 |
| **Torso** | 11.0 | 88 |
| **Mean Detection Accuracy rate = 90.91%** | | |

*Observations*: In Table 1, it can be observed that the head and feet of the proposed system were more properly identifiable because the head was often at the top of the silhouette and the feet were at the bottom of the silhouette. These parts of the body were easier to detect than the other parts like the hips and knees which are in more complex relationships with the body parts, especially when in motion.

### *4.3. Experimentation II: Activity Recognition Accuracies*

For calculating action recognition accuracies, the proposed system was examined by the leave-one-out (LOO) cross-validation method for training and testing data. Table 2 presents the confusion matrix of PAR over the UCF YouTube action dataset and Table 3 represents recognition accuracies of the IM-DailyRGBEvents dataset.

**Table 2.** Confusion matrix of proposed method for UCF YouTube action dataset.

| Activities | SW | DV | TJ | WK | BB | VB | SJ | GS | HR | BK | TS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **SW** | **96.20** | 2.00 | 0.30 | 0 | 0.50 | 0 | 0 | 0 | 0 | 0 | 1.0 |
| **DV** | 0 | **94.70** | 0.33 | 0.58 | 0.10 | 0.55 | 0.46 | 1.98 | 0.32 | 0.12 | 0.86 |
| **TJ** | 0.10 | 1.02 | **89.09** | 3.15 | 0.37 | 1.10 | 0.70 | 0.26 | 1.15 | 1.99 | 1.07 |
| **WK** | 0.20 | 1.92 | 3.07 | **83.40** | 0.91 | 0 | 2.36 | 0.81 | 0.95 | 0.18 | 6.20 |
| **BB** | 5.50 | 1.60 | 2.21 | 2.49 | **86.80** | 0 | 0 | 1.40 | 0 | 0 | 0 |
| **VB** | 0 | 1.41 | 0.81 | 4.71 | 2.74 | **90.20** | 0.05 | 0 | 0.06 | 0 | 0.02 |
| **SJ** | 0.50 | 1.02 | 1.82 | 6.43 | 0 | 0.09 | **88.20** | 0.74 | 0.90 | 0 | 0.30 |
| **GS** | 0 | 1.65 | 0.16 | 0.46 | 3.90 | 0 | 0.80 | **85.80** | 0 | 7.23 | 0 |
| **HR** | 0 | 0.17 | 2.44 | 2.40 | 1.09 | 0.03 | 0 | 4.53 | **87.0** | 2.34 | 0 |
| **BK** | 0 | 0.57 | 0.45 | 0.34 | 0 | 0 | 2.0 | 5.21 | 0.71 | **88.14** | 2.58 |
| **TS** | 0.50 | 0.51 | 0.31 | 0.54 | 0.58 | 0.03 | 0.43 | 0 | 6.61 | 0 | **90.49** |
| **Mean Recognition Accuracy = 89.09%** | | | | | | | | | | | |

Swing = SW; diving = DV; t-jumping = TJ; walking = WK; basketball = BB; volleyball = VB; soccer juggling = SJ; g-swing = GS; horse-riding = HR; biking = BK and t-swing = TS.

**Table 3.** Confusion matrix of the proposed method for the IM-DailyRGBEvents dataset.

| Activities | BX | CP | TO | TH | RA | PC | CN | KG | ET | SD | BD | RW | BW | EX | SU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **BX** | **87.2** | 0 | 0 | 3.0 | 0.05 | 0.13 | 1.29 | 2.0 | 0 | 2.93 | 0 | 0 | 0 | 3.0 | 0.40 |
| **CP** | 0 | **95.4** | 0 | 0.2 | 0 | 0 | 0 | 0 | 4.0 | 0 | 0 | 0 | 0 | 0 | 0.4 |
| **TO** | 0 | 2.31 | **91.0** | 0 | 1.0 | 0 | 0.38 | 2.31 | 0 | 2.0 | 0 | 0 | 1.0 | 0 | 0 |
| **TH** | 0 | 0 | 0 | **91.3** | 0 | 0 | 0 | 0 | 1.0 | 5.7 | 0 | 0.38 | 0 | 1.62 | 0 |
| **RA** | 0 | 0 | 2.0 | 11.0 | **85.8** | 0 | 0 | 0 | 0 | 0 | 1.20 | 0 | 0 | 0 | 0 |
| **PC** | 0 | 10.0 | 0 | 0 | 0.2 | **89.0** | 0.8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **CN** | 0 | 0 | 0 | 0 | 0 | 8.0 | **88.9** | 0 | 0 | 0 | 0 | 0.02 | 1.1 | 1.98 | 0 |
| **KG** | 0 | 0 | 2.0 | 6.5 | 0 | 0 | 0.5 | **86.3** | 0 | 0 | 0 | 0 | 3.9 | 0.80 | 0 |
| **ET** | 0 | 0 | 0 | 0 | 0 | 2.0 | 0 | 6.2 | **88.8** | 3.0 | 0 | 0 | 0 | 0 | 0 |
| **SD** | 0 | 3.0 | 0 | 0 | 5.85 | 0.2 | 0 | 0 | 0 | **90.8** | 0 | 0 | 0 | 0 | 0.15 |
| **BD** | 2.0 | 2.09 | 0 | 2.0 | 0 | 0.31 | 0 | 0 | 0 | 0 | **93.6** | 0 | 0 | 0 | 0 |
| **RW** | 1.2 | 0 | 0 | 0 | 3.10 | 10.0 | 0 | 0.7 | 0 | 0.4 | 0 | **84.6** | 0 | 0 | 0 |
| **BW** | 0 | 0 | 3.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.4 | 18.0 | **77.0** | 0 | 0.6 |
| **EX** | 1.6 | 0 | 0 | 3.0 | 0 | 0.36 | 5.0 | 0.77 | 0.2 | 0.17 | 0 | 0 | 0 | **88.9** | 0 |
| **SU** | 0 | 0.20 | 0 | 0 | 1..0 | 0 | 3.13 | 3.72 | 0 | 0 | 0 | 0 | 0 | 6.7 | **85.25** |
| **Mean Recognition Accuracy = 88.26%** | | | | | | | | | | | | | | | |

Boxing = BX; clapping = CP; take an object = TO; throwing = TH; reading an article = RA; phone conversation = PC; cleaning = CN; kicking = KG; eating = ET; sitting down = SD; bending = BD; right hand waving = RW; both hands waving = BW; exercise = EX and standing up = SU.

*Observations*: In Table 2, it is observed that a few activities such as walking and G-swing affected accuracy due to similarities in patterns with other activities. However, the overall confusion matrix shows significant results of 89.09%. In Table 3, clapping activity shows higher recognition accuracy as it is an easily differentiable activity. On the other hand, recognition accuracy for both hand waving

and Right hand waving was relatively low due to similar motions in these activities. The mean of recognition accuracy scores for the IM-DailyRGBEvents dataset was 88.26%.

### *4.4. Experimentation III: Comparison of the Proposed System with Well-Known Machine Learning Algorithms*

In the third experiment, the results of our proposed system were compared with the results of more commonly used machine learning algorithms. The first algorithm which was chosen for comparison is support vector machine (SVM) and the second algorithm was decision tree. For body parts detection and activity recognition, convolutional neural network (CNN) has gained much popularity due to its effectiveness, so we also selected this algorithm for comparison of the results. In Figure 11, body parts detection results were compared with common machine learning techniques using the UCF sports action dataset. The proposed method's accuracy was 90.91% which was better than CNN's 83%, decision tree's 80% and SVM's 78%. Figures 12 and 13 illustrate the activities recognition results for the UCF YouTube action dataset and the IM-DailyRGBEvents dataset, respectively.

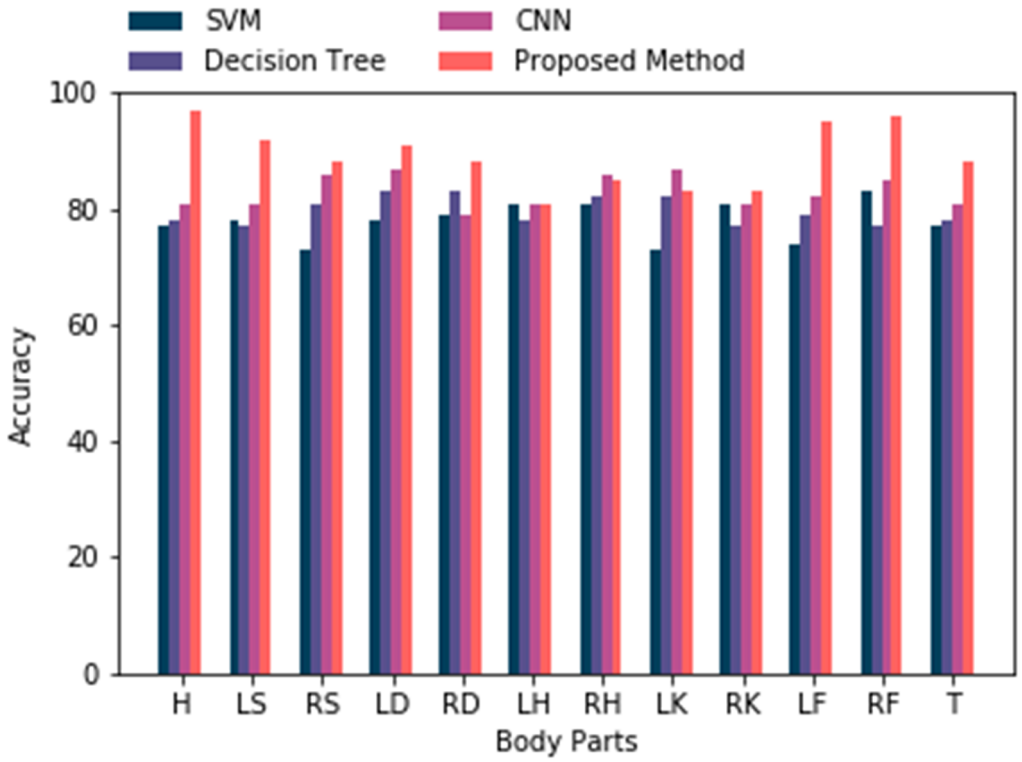



**Figure 11.** Comparison of body parts detection accuracies with common machine learning algorithms on the UCF sports action dataset.

For the UCF YouTube action dataset, the proposed method's accuracy was 89.09% which was better than CNN's 83%, decision tree's 79% and SVM's 80%. For the IM-DailyRGBEvents dataset, the proposed method's accuracy was 88.26% which is better than CNN's 84%, decision tree's 81% and SVM's 77%.

*Observations:* In Figure 11, it can be observed that our method performs better than the other techniques. The performance of CNN is slightly below our method but in some cases detection accuracy of SVM is better than decision tree and CNN. Similarly, in the case of activity recognition as shown in Figures 12 and 13, our method has the best result. However, in a few cases i.e., *both hands waving* in the IM-DailyRGBEvents dataset, accuracy rates of decision tree and of the CNN were slightly better than the accuracy rate of our proposed technique. Similarly, for *reading an article* in the IM-DailyRGBEvents dataset, the accuracy rate of our proposed system was a little below the accuracy

rate of CNN. In conclusion, the overall accuracies of our proposed system for body part detection as well as for action recognition were satisfactory.

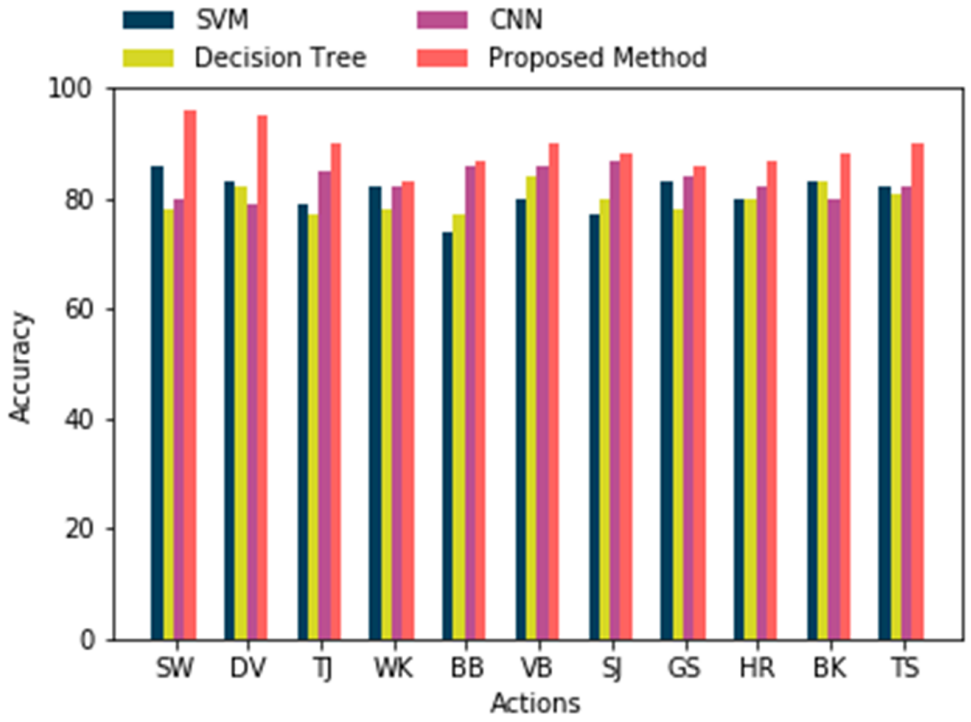



**Figure 12.** Comparison of activity recognition accuracies with common machine learning algorithms on the UCF YouTube action dataset.

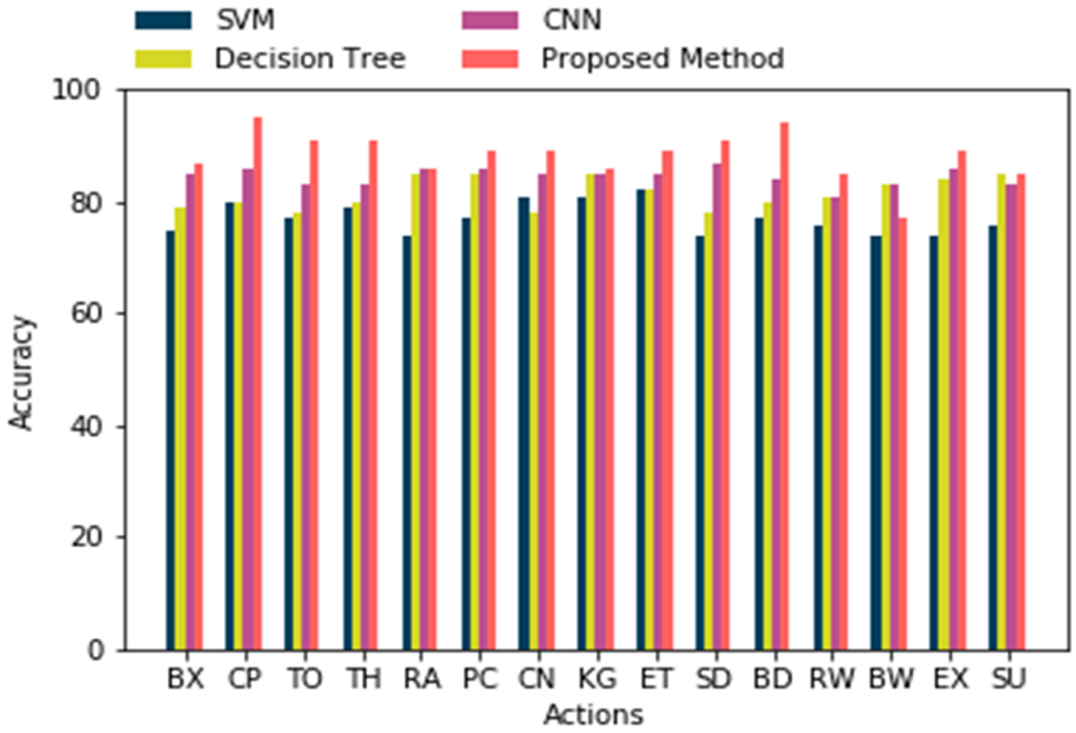



**Figure 13.** Comparison of activity recognition accuracies with common machine learning algorithms on the IM-DailyRGBEvents dataset.

### *4.5. Experimentation IV: Comparison of our Proposed System with State-of-the-Art Techniques*

Table 4 compares the body parts detection accuracy of the proposed multidimensional features method with other state-of-the-art methods using the UCF sports action dataset. It was observed that the proposed method achieved a better detection accuracy rate of 90.91% compared to the others.

**Table 4.** Comparison of the proposed body parts detection accuracies method with other state-of-the-art methods.

| Methods | UCF Sports Actions Dataset (%) |
|---|---|
| **Physical Sports Movements [27]** | 86.67 |
| **HOIRM feature fusion [28]** | 88.25 |
| **Hybrid deep learning model [29]** | 89.01 |
| **Proposed method** | **90.91** |

A comparison of overall results shows that the proposed method achieved a significant improvement with recognition results as high as 89.09% and 88.26% over other methods as shown in Table 5.

**Table 5.** Result comparison of the state-of-the-art methods with proposed physical activity recognition (PAR) method.

| Methods | UCF YouTube Actions (%) | IM-DailyRGBEvents (%) |
|---|---|---|
| **HOJ3D [30]** | 75.5 | - |
| **3D-TCCHOGAC [31]** | 80.27 | - |
| **CSF + TSI-MKL [32]** | 87.81 | - |
| **Proposed method** | **89.09** | **88.26** |

## 5. Conclusions

We proposed a novel technique that combines multidimensional features along with MEMM to detect daily life-log activities for smart indoor/outdoor environments. These features were extracted by robust body part models having 12 tracked key points with an overall accuracy of 90.91%. Finally, the QDA and Markov models were used for optimal discrimination and efficient classification of the extracted features. Experimental results revealed impressive performance (89.09% accuracy for the YouTube action dataset and 88.26% accuracy for the IM-DailyRGBEvents dataset) for the proposed technique and demonstrated that MEMM were used for successful recognition modelling. In future work, we will apply our work to local hospital, fitness gymnasium and kindergarten environments to increase the experimental data sets and make the proposed model more universally applicable.

**Author Contributions:** Conceptualization, A.N.; methodology, A.N. and A.J.; software, A.N.; validation, A.J.; formal analysis, K.K.; resources, A.J. and K.K.; writing—review and editing, A.J. and K.K.; funding acquisition, A.J. and K.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (No. 2018R1D1A1A02085645).

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

## References

1. Trong, N.P.; Minh, A.T.; Nguyen, H.V.; Kazunori, K.; Hoai, B.L. A survey about view-invariant physical activity recognition. In Proceedings of the 2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), Kanazawa University, Kanazawa, Japan, 19–22 September 2017.
2. Shokri, M.; Tavakoli, K. A review on the artificial neural network approach to analysis and prediction of seismic damage in infrastructure. *Int. J. Hydromechatronics* **2019**, *4*, 178–196. [CrossRef]

3. Osterland, S.; Weber, J. Analytical analysis of single-stage pressure relief valves. *Int. J. Hydromechatronics* **2019**, *2*, 32–53. [CrossRef]
4. Jalal, A.; Kim, Y. Dense Depth Maps-based Human Pose Tracking and Recognition in Dynamic Scenes Using Ridge Data. In Proceedings of the 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Seoul, Korea, 26–29 August 2014.
5. Trung, N.T.; Makihara, Y.; Nagahara, H.; Mukaigava, Y.; Yagi, Y. Inertial-sensor-based walking action recognition using robust step detection and inter-class relationships. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012.
6. Trung, N.T.; Makihara, Y.; Nagahara, H.; Mukaigava, Y.; Yagi, Y. Similar gait action recognition using an inertial sensor. *Pattern Recognit.* **2015**, *48*, 1289–1301.
7. Hawang, I.; Cha, G.; Oh, S. Multi-modal human action recognition using deep neural networks fusing image and inertial sensor data. In Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Daegu, Korea, 16–18 November 2017.
8. Irvin, L.-N.; Muñoz-Meléndez, A. Human action recognition based on low- and high-level data from wearable inertial sensors. *Int. J. Distrib. Sens. Netw.* **2019**, *15*, 1–12.
9. Dawar, N.; Ostadabbas, S.; Kehtarnavaz, N. Data Augmentation in Deep Learning-Based Fusion of Depth and Inertial Sensing for Action Recognition. *IEEE Sens. Lett.* **2019**, *3*, 1–4. [CrossRef]
10. Fang, H.; Thiyagalingam, J.; Bessis, N.; Edirisinghe, E. Fast and reliable human action recognition in video sequences by sequential analysis. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017.
11. Silambarasi, R.; Sahoo, S.P.; Ari, S. 3D spatial-temporal view based motion tracing in human action recognition. In Proceedings of the 2017 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 6–8 April 2017.
12. Shehzed, A.; Jalal, A.; Kim, K. Multi-Person Tracking in Smart Surveillance System for Crowd Counting and Normal/Abnormal Events Detection. In Proceedings of the 2019 International Conference on Applied and Engineering Mathematics (ICAEM), Taxila, Pakistan, 27–29 August 2019.
13. Han, Y.; Chung, S.L.; Ambikapathi, A.; Chan, J.S.; Lin, W.Y.; Su, S.F. Robust human action recognition using global spatial-temporal attention for human skeleton data. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018.
14. Susan, S.; Agrawal, P.; Mittal, M.; Bansal, S. New shape descriptor in the context of edge continuity. *CAAI Trans. Intell. Technol.* **2019**, *4*, 101–109. [CrossRef]
15. Dwina, N.; Arnia, F.; Munadi, K. Skin segmentation based on improved thresholding method. In Proceedings of the 2018 International ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI-NCON), Chiang Rai, Thailand, 25–28 February 2018.
16. Gomathi, S.; Santhanam, T. Application of Rectangular Feature for Detection of Parts of Human Body. *Adv. Comput. Sci. Technol.* **2018**, *11*, 43–55.
17. Tingting, Y.; Junqian, W.; Lintai, W.; Yong, X. Three-stage network for age estimation. *CAAI Trans. Intell. Technol.* **2019**, *4*, 122–126. [CrossRef]
18. Wiens, T. Engine speed reduction for hydraulic machinery using predictive algorithms. *Int. J. Hydromechatronics* **2019**, *1*, 16–31. [CrossRef]
19. Yao, L.; Min, W.; Lu, K. A new approach to fall detection based on the human torso motion model. *Appl. Sci.* **2017**, *7*, 993. [CrossRef]
20. Matsukawa, T.; Suzuki, E. Kernelized cross-view quadratic discriminant analysis for person re-identification. In Proceedings of the 2019 16th International Conference on Machine Vision Applications (MVA), Tokyo, Japan, 27–31 May 2019.
21. Zhu, C.; Miao, D. Influence of kernel clustering on an RBFN. *CAAI Trans. Intell. Technol.* **2019**, *4*, 255–260. [CrossRef]
22. Wang, H.; Fei, H.; Yu, Q.; Zhao, W.; Yan, J.; Hong, T. A motifs-based Maximum Entropy Markov Model for realtime reliability prediction in System of Systems. *J. Syst. Softw.* **2019**, *151*, 180–193. [CrossRef]
23. Nuruzzaman, M.; Hussain, O.K. Identifying facts for chatbot's question answering via sequence labelling using recurrent neural networks. In Proceedings of the ACM Turing Celebration Conference—China, Chengdu, China, 17–19 May 2019.

24. Rodriguez, M.D.; Ahmed, J.; Shah, M. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 24–26 June 2008.
25. Liu, J.; Luo, J.; Shah, M. Recognizing Realistic Actions from Videos "in the Wild". In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
26. Jalal, A.; Uddin, M.Z.; Kim, T.S. Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home. *IEEE Trans. Consum. Electron.* **2012**, *58*, 863–871. [CrossRef]
27. Jalal, A.; Nadeem, A.; Bobasu, S. Human Body Parts Estimation and Detection for Physical Sports Movements. In Proceedings of the 2nd International Conference on Communication, Computing and Digital systems (C-CODE), Islamabad, Pakistan, 6–7 March 2019.
28. Huan, R.H.; Xie, C.J.; Guo, F.; Chi, K.K.; Mao, K.J.; Li, Y.L.; Pan, Y. Human action recognition based on HOIRM feature fusion and AP clustering BOW. *PLoS ONE* **2019**, *14*, 1–15. [CrossRef] [PubMed]
29. Jaouedi, N.; Boujnah, N.; Bouhlelc, M.S. A new hybrid deep learning model for human action recognition. *J. King Saud Univ. Comput. Inf. Sci.* **2020**, *32*, 447–453. [CrossRef]
30. Li, N.; Cheng, X.; Zhang, S.; Wu, Z. Realistic human action recognition by Fast HOG3D and self-organization feature map. *Mach. Vis. Appl. Vol.* **2014**, *25*, 1793–1812. [CrossRef]
31. Tong, M.; Wang, H.; Tian, W.; Yang, S. Action recognition new framework with robust 3D-TCCHOGAC and 3D-HOOFGAC. *Multimed. Tools Appl.* **2017**, *76*, 3011–3030. [CrossRef]
32. Yang, Y.; Hu, P.; Deng, X. Human action recognition with salient trajectories and multiple kernel learning. *Multimed. Tools Appl.* **2018**, *77*, 17709–17730.