applied sciences

MDPI

Article

Human Action Recognition from Multiple Views Based on View-Invariant Feature Descriptor Using Support Vector Machines

Allah Bux Sargano [1,2,*], Plamen Angelov [1] and Zulfiqar Habib [2]

[1] School of Computing and Communications Infolab21, Lancaster University, Lancaster LA1 4WA, UK; p.angelov@lancaster.ac.uk

[2] Department of Computer Science, COMSATS Institute of Information Technology, Lahore 54000, Pakistan; drzhabib@ciitlahore.edu.pk

* Correspondence: a.bux@lancaster.ac.uk; Tel.:+44-152-451-0525

Academic Editor: David He

Received: 5 September 2016; Accepted: 13 October 2016; Published: 21 October 2016

Abstract: This paper presents a novel feature descriptor for multiview human action recognition. This descriptor employs the region-based features extracted from the human silhouette. To achieve this, the human silhouette is divided into regions in a radial fashion with the interval of a certain degree, and then region-based geometrical and Hu-moments features are obtained from each radial bin to articulate the feature descriptor. A multiclass support vector machine classifier is used for action classification. The proposed approach is quite simple and achieves state-of-the-art results without compromising the efficiency of the recognition process. Our contribution is two-fold. Firstly, our approach achieves high recognition accuracy with simple silhouette-based representation. Secondly, the average testing time for our approach is 34 frames per second, which is much higher than the existing methods and shows its suitability for real-time applications. The extensive experiments on a well-known multiview IXMAS (INRIA Xmas Motion Acquisition Sequences) dataset confirmed the superior performance of our method as compared to similar state-of-the-art methods.

Keywords: computer visions; human action recognition; view-invariant feature descriptor; classification; support vector machines

1. Introduction

In recent years, automatic human-activity recognition (HAR) based on computer vision has drawn much attention of researchers around the globe due to its promising results. The major applications of HAR include; human–computer interaction (HCI), intelligent video surveillance, ambient assisted living, human–robot interaction, entertainment, video indexing, and others [1]. Depending on the complexity and duration, human activities can be categorized into four levels: gestures, actions, interactions, and group activities. Gestures are represented by movement of the person's body parts, such as stretching an arm; actions are single-person activities such as walking, punching, kicking, running and so forth; interactions are activities of two or more than two persons, such as two person fighting with each other; and a group having a meeting is an example of group activity [2]. A large amount of work has already been done for human action recognition, but it is still a challenging problem.

The major challenges and issues in HAR are as follows: (1) occlusion; (2) variation in human appearance, shape, and clothes; (3) cluttered backgrounds; (4) stationary or moving cameras; (5) different illumination conditions; and (6) viewpoint variations. Among these challenges, viewpoint variation is one of the major problems in HAR since most of the approaches for human activity

classification are view-dependent and can recognize the activity from one fixed view captured by a single camera. These approaches are supposed to have the same camera view during training and testing. This condition cannot be maintained in real world application scenarios. Moreover, if this condition is not met, their accuracy decreases drastically because the same actions look quite different when captured from different viewpoints [3]. A single camera-based approach also fails to recognize the action when an actor is occluded by an object or when some parts of the action are hidden due to unavoidable self-occlusion. To avoid these issues and get the complete picture of an action, more than one camera is used to capture the action—this is known as action recognition from multiple views or view-invariant action recognition [4].

There are two major approaches for recognition of human actions from multiviews: 3D approach and 2D approach [5]. In the first approach, a 3D model of a human body is constructed from multiple views and motion representation is formed from it for action recognition. This model can be based on cylinders, ellipsoids, visual hulls generated from silhouettes, or surface mesh. Some examples of motion representation are 3D optical flow [6], shape histogram [7], motion history volume [8], 3D body skeleton [9], and spatiotemporal motion patterns [10]. Usually, the 3D approach provides higher accuracy than a 2D approach but at higher computational cost, which makes it less applicable for real time applications. In addition to this, it is difficult to reconstruct a good-quality 3D model because it depends on the quality of extracted features or silhouettes of different views. Hence, the model is exposed to deficiencies which might have occurred due to segmentation errors in each viewpoint. Moreover, a good 3D model of different views can only be constructed when the views overlap. Therefore, a sufficient number of viewpoints have to be available to reconstruct a 3D model.

However, recently some 3D cameras have been introduced for capturing images in 3D form. Among these, 3D time-of-flight (ToF) cameras and Microsoft Kinect have become very popular for 3D imaging. These devices overcome the difficulties that are faced by the classical 3D multiview action-recognition approaches when reconstructing a 3D model. However, these sensors also have several limitations. For example, in contrast to a fully reconstructed 3D model from multiple views, these sensors only capture the frontal surfaces of the human and other objects in the scene. In addition to this, these sensors also have limited range about 6–7 m, and data can be distorted by scattered light from the reflective surfaces [5]. Due to the limitations of the 3D approach, researchers prefer to employ a 2D approach for human-action recognition from multiple views [11].

The methods based on 2D models extract features from 2D images covering multiple views. Different methods have been proposed for multiview action recognition based on 2D models. However, three important lines of work are mentioned here. The first approach handles it at feature level, achieves view-invariant action representation using appropriate feature descriptor(s) or fusion of different features [12,13], and then action recognition is performed using an appropriate classifier. The second one handles it at a classification level by determining the appropriate classification scheme. The classification is carried out either by a single universal classifier or multiple classifiers are trained, and later on their results are fused to get the final result [14,15]. The third one utilizes the learning-based model, such as deep learning and its variations, to learn the effective and discriminative features directly from the raw data for multiview action recognition [16,17].

Our method falls under the first category: we learn the discriminative and view-invariant features from the human silhouette. The human silhouette contains much less information than the original image, but our approach shows that this information is sufficient for action representation and recognition. By taking advantage of this simple representation, our approach not only provides high accuracy but also high recognition speed of 34 frames per second on a challenging multiview IXMAS (INRIA Xmas Motion Acquisition Sequences) dataset. Due to its high accuracy and speed our approach is suitable for real-time applications. The rest of the paper is organized as follows. Related work is discussed in Section 2, proposed approach is described in Section 3, experimentation and results are explained in Section 4, finally discussions and conclusion are presented in Section 5.

2. Related Works

This section presents state-of-the-art methods for multiview action recognition based on a 2D approach. These methods extract features from 2D image frames of all available views and combine these features for action recognition. Then, classifier is trained using all these viewpoints. After training the classifier, some methods use all viewpoints for classification [18], while others use a single viewpoint for classification of a query action [19–21]. In both cases, the query view is part of the training data. However, if the query view is different than the learned views, this is known as cross-view action recognition. This is even more challenging than the multiview action recognition [22,23].

Different types of features—such as motion features, shape features, or combination of motion- and shape-based features—have been used for multiview action recognition. In [20], silhouette-based features were acquired from five synchronized and calibrated cameras. The action recognition from multiple views was performed by computing the R transform of the silhouette surfaces and manifold learning. In [24], contour points of the human silhouette were used for pose representation, and multiview action recognition was achieved by the arrangements of multiview key poses. Another silhouette-based method was proposed in [25] for action recognition from multiple views; this method used contour points of the silhouette and radial scheme for pose representation. Then, model fusion of multiple camera streams was used to build the bag of key poses, which worked as a dictionary for known poses and helped to convert training sequences into key poses for a sequence-matching algorithm. In [13], a view-invariant recognition method was proposed, which extracted the uniform rotation-invariant local binary patterns (LBP) and contour-based pose features from the silhouette. The classification was performed using a multiclass support vector machine. In [26], scale-invariant features were extracted from the silhouette and clustered to build the key poses. Finally, classification was done using a weighted voting scheme.

An optical flow and silhouette-based features were used for view-invariant action recognition in [27], and principal component analysis (PCA) was used for reducing the dimensionality of the data. In [28], coarse silhouette features, radial grid-based features and motion features were used for multiview action recognition. Another method for viewpoint changes and occlusion-handling was proposed in [19]. This method used histogram of oriented gradients (HOG) features with local partitioning, and obtained the final results by fusing the results of the local classifiers. A novel motion descriptor based on motion direction and histogram of motion intensity was proposed in [29] for multiview action recognition followed by a support vector machine used as a classifier. Another method based on 2D motion templates, motion history images, and histogram of oriented gradients was proposed in [30]. A hybrid CNN–HMM model which combines convolution neural networks (CNN) with hidden Markov model (HMM) was used for action classification [17]. In this method, the CNN was used to learn the effective and robust features directly from the raw data, and HMM was used to learn the statistical dependencies over the contiguous subactions and conclude the action sequences.

3. Proposed System

In recent years, various methods have been published for multiview action recognition, but very few are actually suitable for real-time applications due to their high computational cost. Therefore, the cost of the feature extraction and action classification has to be reduced as much as possible. The proposed system has been designed around these parameters. It uses the human silhouette as input for feature extraction and formulating the region-based novel feature descriptor. Human silhouette can easily be extracted using foreground detection techniques. At the moment our focus is not on foreground segmentation; rather, we are focused on later phases of the action recognition such as feature extraction and classification. Although the human silhouette is a representation that contains much less information than the original image, we show that it contains sufficient information to recognize the human actions with high accuracy. This work proposes a novel feature descriptor

for multiview human-action recognition based on human silhouette. When a silhouette is partitioned into radial bins, then region descriptors are formed which are represented as a triangle or quadrangle. Accordingly, it makes the pose description process simple and accurate. The descriptive detail of the proposed system is given in Algorithm 1 and block diagram is shown in Figure 1.

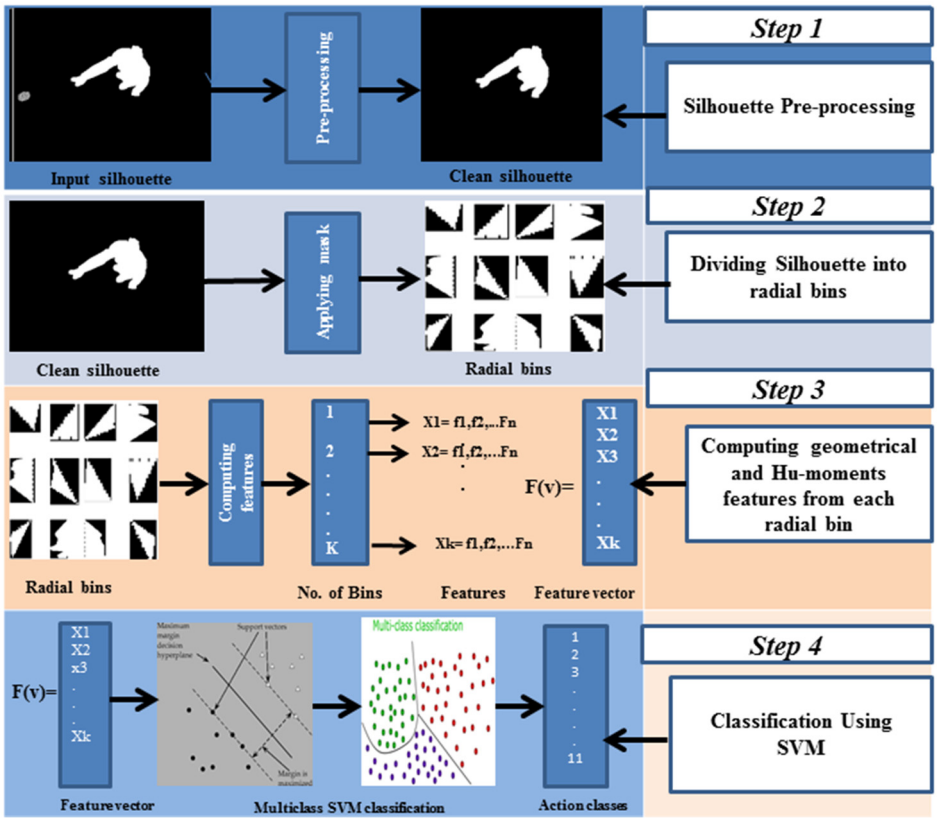

Figure 1. Block diagram of the proposed system.

We consider the centroid of a silhouette as a point of origin, and divide the human silhouette into R radial bins with the interval of same degree. Here the value of the R is selected as 12, which constitutes 12 radial bins with the interval of 30° each. This value has been selected with experimentation to cover the maximum viewing angles. This is unlike [31]—where the radial histograms were used as a feature descriptor—and [25], where the contour points of each radial bin were considered as features, and model fusion was used to achieve the multiview action recognition by obtaining key poses for each action through K-means clustering. Our method computes efficient and discriminative geometrical and Hu-moment features from each radial bin of the silhouette itself. Then, these features from all bins are concatenated into a feature vector for multiview action recognition using support vector machine.

This approach has three major advantages. Firstly, it divides the silhouettes into radial bins, which cover almost all viewing angles, thus provides an easy way to compute the features for different views of an action. Secondly, it uses the selected discriminative features and avoids extra computation such as fusion or clustering as used in [25]. Thirdly, due to employing selected features from each bin, it does not require any dimensionality reduction technique.

Algorithm 1

Step 1 Input video from multiple cameras
Step 2 Extraction of silhouettes from the captured video
Step 3 Feature Extractions:

a) Division of the human silhouette into radial bins
b) Computation of region-based geometrical features from each radial bin
c) Computation of Hu-moments features from each radial bin

Step 4 Action classifications by multi-class support vector machine
Step 5 Output recognized action

3.1. Pre-Processing

Usually, some lines or small size regions are formed around the silhouette due to segmentation errors, loose clothing of the subject under consideration, and other noise. These unnecessary regions do not offer any important information for action recognition, but rather create problems for the feature-extraction algorithm and increase the complexity. By removing these small regions, complexity can be reduced and the feature-extraction process can be made more accurate. In our case, a region was considered as small and unnecessary if its area is less than 1/10 of the silhouette. This threshold was set based on an observation on 1000 silhouette images of different actors and actions. An example of the silhouette before and after noise removal is shown in first row of Figure 1.

3.2. Multiview Features Extraction and Representation

The success of any recognition system mainly depends on proper feature selection and extraction mechanism. For action recognition from different views, a set of discriminative and view-invariant features have to be extracted. Our feature descriptor is based on two types of features: (1) region-based geometric features and (2) Hu-moments features extracted from each radial bin of the silhouette. The overview of the feature extraction process is show in Figure 2.

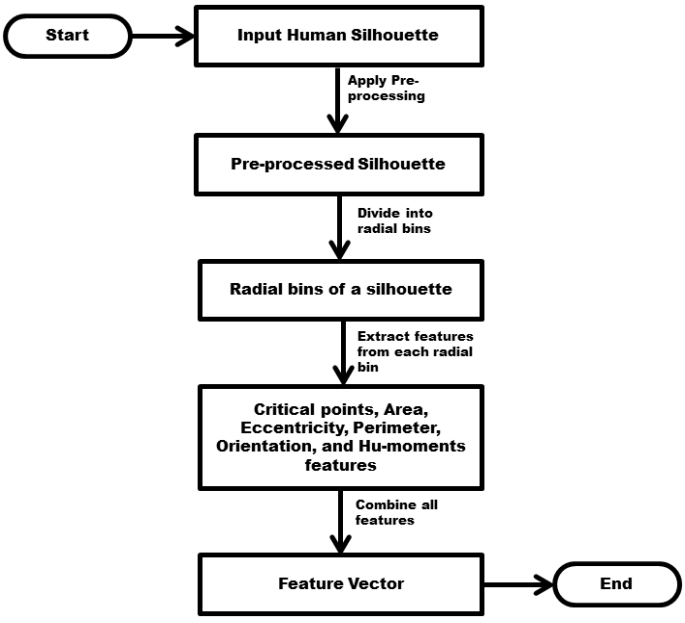

Figure 2. Overview of feature extraction process.

3.2.1. Region-Based Geometric Features

The shape of a region can be described by two types of features. One is the region-based and the other is boundary-based features. The region-based features are less affected by noise and occlusion than boundary-based features. Therefore, region-based features are better choice to describe the shape of a region [32]. Generally, a shape is described by a set of numbers known as descriptors. A good descriptor is one which has the ability to reconstruct the shape from the feature points. The resultant shape should be an approximation of the original shape and should yield similar feature values. The proposed feature descriptor has been designed around these parameters. These features have been computed as follows:

(1) First of all, we calculate the centroid of a human silhouette using Equation (2), as shown in Figure 3.

$$C_m = (x_c,\ y_c) \tag{1}$$

where

$$x_c = \frac{\sum_{i=1}^{n} x_i}{n} \text{ and } y_c = \frac{\sum_{i=1}^{n} y_i}{n} \tag{2}$$

(2) After computing the centroid by Equation (2), a radius of the silhouette is computed.

(3) The silhouette is divided into 12 radial bins with respect to its centroid using a mask; this division has been made with intervals of 30°. These bins are shown in Figure 4.

(4) The following region-based features are computed for each bin of the silhouette.

(a) *Critical points:* As we move the mask on the silhouette in counter-clockwise direction, a triangle or quadrangle shape is formed in each bin. We compute the critical points (corner points) of each shape and their distances. There can be different numbers of critical points for each shape; therefore, the mean and variance of these points have been computed as features. It gives us $2 \times 12 = 24$ features for each silhouette.

(b) *Area:* The simple and natural property of a region is its area. In the case of a binary image, it is a measure of size of its foreground. We have computed the area of each bin, which provides 12 important features for each human silhouette.

(c) *Eccentricity:* The ratio of the major and minor axes of an object is known as eccentricity [33]. In our case, it is ratio of distance between the major axis and foci of the ellipse. Its value is between 0 and 1, depending upon the shape of the ellipse. If its value is 0, then actually it is a circle; if its value is 1, then it is a line segment. We have computed eccentricity for each of the 12 bins.

(d) *Perimeter:* This is also an important property of a region. The distance around the boundary of a region can be measured by computing the distance between each pair of pixels. We have computed perimeter of each bin forming a triangle or quadrangle.

(e) *Orientation:* This is an important property of a region, which specifies the angle between x-axis and major axis of the ellipse. Its value can be between $-90°$ and $90°$.

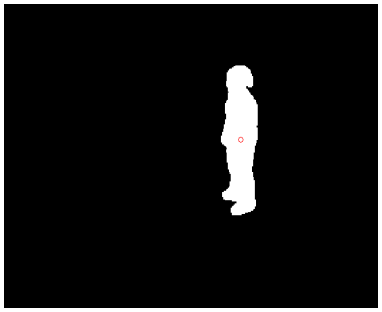

Figure 3. Example image of the human silhouette with centroid.

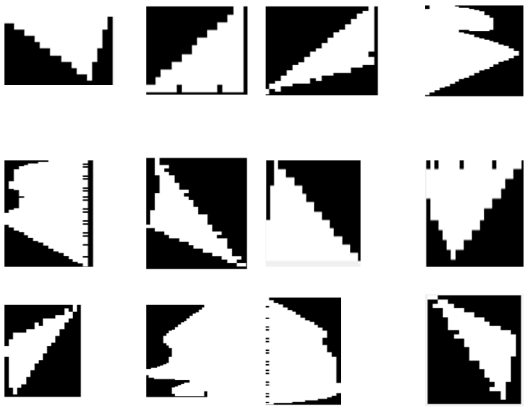

Figure 4. Example of division of silhouette into radial bins.

3.2.2. Hu-Moments Invariant Features

The use of invariant moments for binary shape representation was proposed in [34]. The moments which are invariant with respect to rotation, scales, and translations, are known as Hu-moments invariants. We have computed seven moments for each radial bin of the silhouette. The moment (p,q) of an image $f(x,y)$ of size $M \times N$ is defined as:

$$m_{p,q} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) x^p y^q \tag{3}$$

Here, p is order of x and q is order of y. We can calculate the central moment in the same way as these moments, except the value of x and y is displaced by the mean values as follows:

$$\mu_{p,q} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y)(x - x_{avg})^p (y - y_{avg})^q \tag{4}$$

where

$$x_{avg} = \frac{m_{10}}{m_{00}} \text{ and } y_{avg} = \frac{m_{01}}{m_{00}} \tag{5}$$

By applying normalization, scale-invariant moments are obtained. Hence, normalized central moments are defined as follows [35].

$$\eta_{p,q} = \frac{\mu_{p,q}}{\mu 00^{\gamma}}, \ \gamma = \frac{p+q+2}{2}, \ p+q = 2,3,\ldots \tag{6}$$

Based on these central moments, [34] introduced seven Hu-moments as linear combination of central moments defined as follows:

$$\begin{aligned}
h_1 &= \eta_{20} + \eta_{02} \\
h_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta^2{}_{11} \\
h_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
h_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
h_5 &= (\eta_{30} - 3\eta_{12})((\eta_{30} + 3\eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2) + \\
&(3\eta_{21} - \eta_{03})^2(\eta_{21} + \eta_{03})(3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2) \\
h_6 &= (\eta_{20} - \eta_{02})((\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2) + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
h_7 &= (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})(3(\eta_{30} - 3\eta_{12})^2 - (\eta_{21} + \eta_{03})^2) - \\
&(\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})(3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2)
\end{aligned}$$

3.3. Action Classification with SVM Multiclass Classifier

After computing the features from video, an appropriate classifier has to be used for classification of actions. In supervised learning, discriminative models are more effective than generative models [36]. Support vector machine (SVM) is one of the most successful classifiers in discriminative models category. It has shown better performance than some old-style classifiers such as backpropagation neural networks, Naïve Bayes, and k-nearest neighbors (KNNs) in many classification problems [37]. SVM was first proposed in [38], and, originally, it was developed for binary classification but later on extended to the multiclass classification problem. There are two main approaches for multiclass SVM. The first approach considers all classes of data directly into one optimization formulation, while the second approach constructs and combines binary classifiers in some manners to build a multiclass classifier. The second approach is computationally less expensive and easy to implement. Many algorithms have been derived for multiclass classification using this approach, such as one-against-all [39], one-against-one [40], Directed Acyclic Graph-Support Vector Machine (DAG-SVM) [41], Error-Correcting Output Codes Support Vector Machine (ECOC-SVM) [42], and Support Vector Machines with Binary Tree Architecture (SVM-BTA) [43]. Among these, one-against-all [39] and one-against-one [40] are two commonly used methods for multiclass classification. The one-against-all needs N SVM binary classifiers for an N class classification problem, while one-against-one method needs $\frac{N(N-1)}{2}$ binary classifiers for an N number of classes, each trained from samples of two corresponding classes. As compared to one-against-all method, one-against-one is better in terms of accuracy for many classification problems [44].

We used multiclass SVM classifier implementation in [45] for multiview action recognition, which uses the one-against-one method with radial basis function (RBF) kernel. Moreover, to estimate the best parameters for classifier, we conducted grid search to know the best value for parameter γ and C. Here, γ represents the width of the RBF kernel and C represents the weight of error penalty. The appropriate set of (C, γ) increases the overall accuracy of the SVM classifier [46].

4. Experimentations

For the evaluation of proposed method, comprehensive experimentations have been conducted on a well-known multiview IXMAS [47] dataset. The leave-one-sequence-out (LOSO) scheme has been used for view-invariance evaluation. In this scheme, the classifier is trained on all sequences except one, which is used for testing. This process is repeated for all possible combinations and results are averaged. This is a common strategy used by different researchers such as [19,29] for evaluation of their methods. This is helpful to compare our results with these state-of-the-art methods.

4.1. Evaluation on Multiview Action Recognition Dataset

The IXMAS is a challenging and well-known dataset with multiple actors and camera views. This dataset is popular among the human-action recognition methods for testing view-invariant action recognition algorithms, including both multiview and cross-view action recognition. It includes 13 daily life action classes with 5 different cameras, including one top-view and four side cameras, as shown in Figure 5. Each action is performed 3 times by 12 different subjects while actors keep changing orientations in each sequence during action execution. The change in orientation is indicated by action labels, and no additional information is provided other than these labels. Most of the existing methods use selected action classes and actors for experimentation [19,29]. For comparison, we have selected 11 action classes performed by 12 actors as shown in Figure 6. The name and the label index of these actions are shown in Table 1.

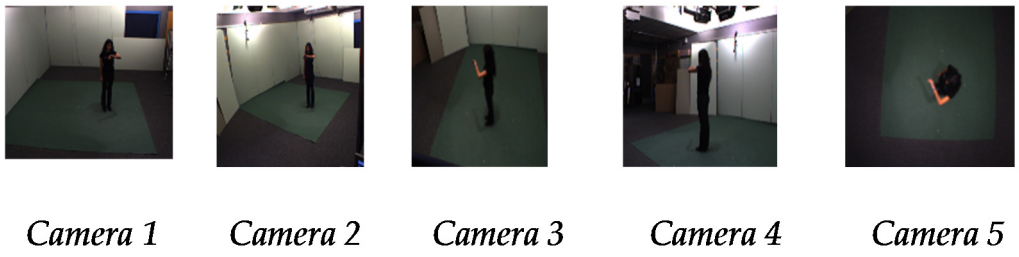

Figure 5. Five cameras views of the same action (check watch).

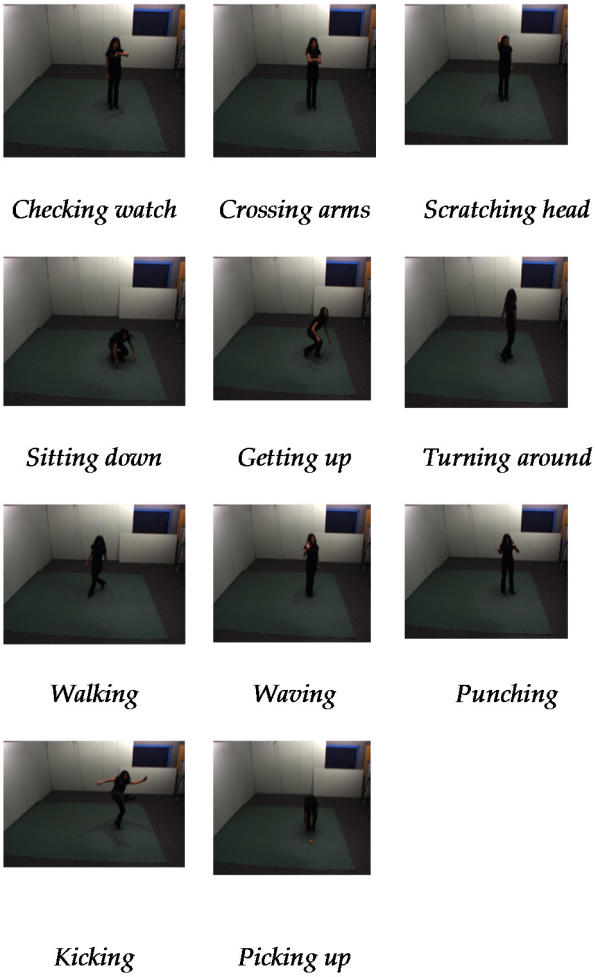

Figure 6. Selected 11 actions in IXMAS dataset.

Table 1. Action class names used in experimentation.

Index	Action Name	Index	Action Name
1	Checking watch	7	Walking
2	Crossing arms	8	Waving
3	Scratching head	9	Punching
4	Sitting down	10	Kicking
5	Getting up	11	Picking up
6	Turning around	-	-

4.2. Comparison with Similar Methods on IXMAS Dataset

Our method achieves a recognition rate of 89.75% with leave-one-sequence-out (LOSO) cross-validation on 11 actions. The recognition rate of individual action is presented in a confusion matrix, shown in Figure 7. The results confirm that our method outperforms the similar state-of-the-art 2D based methods such as [12,19,21,24,29,48–52] recorded in Table 2. It is important to be mentioned here that the number of classes, actors, and views used in experimentations vary among these methods. For example, in [52], 89.4% accuracy has been reported but they excluded camera 4 from experimentation. Likewise, [51] also excluded the top camera and considered only the remaining 4 cameras. Moreover, most of the published methods are not appropriate for real-time application due to their high computational cost. The proposed method considers all views of the IXMAS dataset including the top view for recognition. The results indicate that the proposed method is superior to the similar 2D methods, not only in recognition accuracy but also in recognition speed as well.

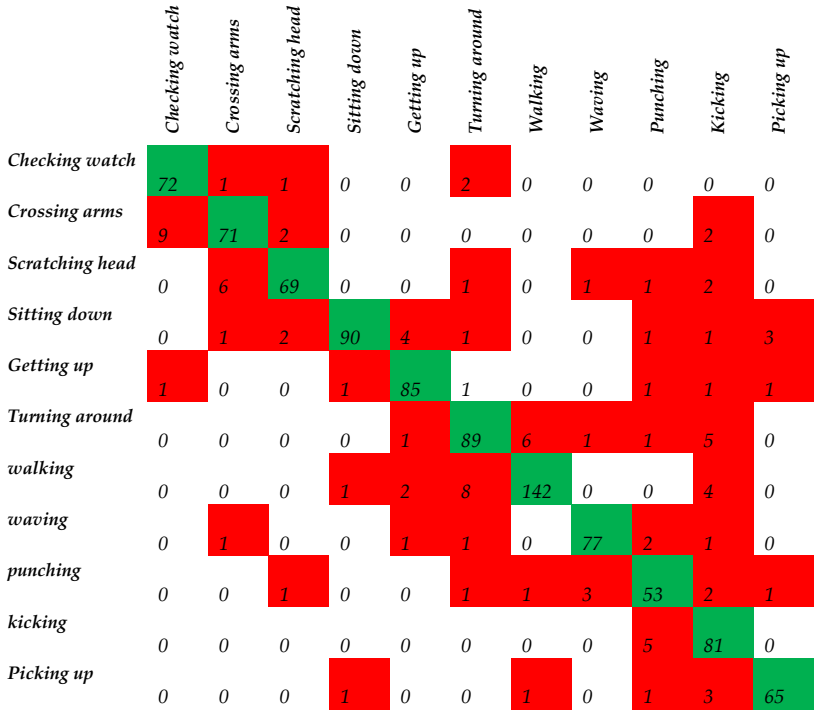

Figure 7. Confusion matrix of IXMAS (INRIA Xmas Motion Acquisition Sequences) dataset with 11 actions.

Table 2. Comparison with state-of-the-art methods on IXMAS (INRIA Xmas Motion Acquisition Sequences) dataset.

Year	Method	Accuracy (%)
-	Proposed method	89.75
2016	Chun et al. [29]	83.03
2013	Chaaraoui et al. [24]	85.9
2013	Burghouts et al. [53]	96.4
2011	Wu et al. [52]	89.4
2011	Junejo et al. [12]	74
2010	Weinland et al. [19]	83.4
2009	Reddy et al. [48]	72.6
2008	Liu and Shah [21]	82.8
2008	Cherla et al. [51]	80.1
2008	Vitaladevuni et al. [50]	87.0
2007	Lv and Nevatia [49]	80.6

The resolution of IXMAS dataset is only 390 × 291, which is very low resolution as compared to many other action recognition datasets. We performed experiments using MATLAB R2015b implementation on Intel® Core i7-4770 CPU with 8 cores @ 3.4 GHz, 8 GB RAM, and Windows 10 operating system. However, only 4 cores of the CPU were utilized during experimentation. Moreover, we did not use any optimization techniques in the code. The average testing time of our approach is 0.0288 per frame, which is almost 34 frames per second (FPS); this is much better than the existing methods as recorded in Table 3.

Table 3. Comparison of average testing speed on IXMAS dataset.

Method	Average FPS	Accuracy (%)
Proposed method	34	89.75
Chaaraoui et al. [24]	26	85.9
Cherla et al. [51]	20	80.1
Lv and Nevatia [49]	5.1	80.6

5. Conclusions

In this paper, a multiview human-action recognition method based on a novel region-based feature descriptor is presented. These methods are divided into two categories: 2D approach- and 3D approach-based methods. Usually, the 3D approach provides better accuracy than a 2D approach, but it is computationally expensive, which makes it less applicable for real-time applications. The proposed method uses the human silhouette as input to the features extraction process. Although the human silhouette contains much less information than the original image, our experimentation confirms that it is sufficient for action recognition with high accuracy. Moreover, to get the view-invariant features, the silhouette is divided into radial bins with the interval of 30° each. Then, carefully selected region-based geometrical features and Hu-moment features are computed for each bin. For action recognition, a multiclass support vector machine with RBF kernel is employed. The proposed method has been evaluated on the well-known IXMAS dataset. This dataset is a challenging benchmark available for evaluation of multiview action recognition methods. The results indicate that our method outperforms the state-of-the-art 2D methods both in terms of efficiency and recognition accuracy. The testing time for our approach is 34 frames second, which makes it very much suitable for real-time applications. As far as more complex datasets such as HMDB-51, YouTube, and Hollywood-II are concerned, we have not tested our approach with these datasets. However, we believe our approach should also produce good accuracy with these datasets as well, because once the silhouette is extracted the scene complexities will not matter very much. However, perfect silhouette extraction in complex scenarios is still a challenging task which may further affect the feature extraction process as well. As a future work we would like to extend our method for cross-view action recognition—a special case of multiview action recognition—where a query view is different than the learned views. We will also evaluate our approach on more complex datasets, as mentioned above.

Acknowledgments: This study is supported by the research grant of COMSATS Institute of Information Technology, Pakistan.

Author Contributions: Allah Bux Sargano, Plamen Angelov, and Zulfiqar Habib conceived and designed the experiments; Allah Bux Sargano performed the experiments; Allah Bux Sargano and Zulfiqar Habib analyzed the data; Allah Bux Sargano and Plamen Angelov contributed reagents/materials/analysis tools; Allah Bux Sargano wrote the paper.

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

References

1. Poppe, R. A survey on vision-based human action recognition. *Image Vis. Comput.* **2010**, *28*, 976–990. [CrossRef]

2. Aggarwal, J.K.; Ryoo, M.S. Human activity analysis: A review. *ACM Comput. Surv. (CSUR)* **2011**, *43*, 16. [CrossRef]
3. Rudoy, D.; Zelnik-Manor, L. Viewpoint selection for human actions. *Int. J. Comput. Vis.* **2012**, *97*, 243–254. [CrossRef]
4. Saghafi, B.; Rajan, D.; Li, W. Efficient 2D viewpoint combination for human action recognition. *Pattern Anal. Appl.* **2016**, *19*, 563–577. [CrossRef]
5. Holte, M.B.; Tran, C.; Trivedi, M.M.; Moeslund, T.B. Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. *IEEE J. Sel. Top. Signal Proc.* **2012**, *6*, 538–552. [CrossRef]
6. Holte, M.B.; Moeslund, T.B.; Nikolaidis, N.; Pitas, I. 3D human action recognition for multi-view camera systems. In Proceedings of the 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), Hangzhou, China, 16–19 May 2011; pp. 324–329.
7. Huang, P.; Hilton, A.; Starck, J. Shape similarity for 3D video sequences of people. *Int. J. Comput. Vis.* **2010**, *89*, 362–381. [CrossRef]
8. Weinland, D.; Ronfard, R.; Boyer, E. Free viewpoint action recognition using motion history volumes. *Comput. Vis. Image Underst.* **2006**, *104*, 249–257. [CrossRef]
9. Slama, R.; Wannous, H.; Daoudi, M.; Srivastava, A. Accurate 3D action recognition using learning on the Grassmann manifold. *Pattern Recognit.* **2015**, *48*, 556–567. [CrossRef]
10. Ali, S.; Shah, M. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 288–303. [CrossRef] [PubMed]
11. Holte, M.B.; Tran, C.; Trivedi, M.M.; Moeslund, T.B. Human action recognition using multiple views: A comparative perspective on recent developments. In Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding, Scottsdale, AZ, USA, 28 November–1 December 2011; pp. 47–52.
12. Junejo, I.N.; Dexter, E.; Laptev, I.; Perez, P. View-independent action recognition from temporal self-similarities. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 172–185. [CrossRef] [PubMed]
13. Kushwaha, A.K.S.; Srivastava, S.; Srivastava, R. Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns. *Multimedia Syst.* **2016**. [CrossRef]
14. Iosifidis, A.; Tefas, A.; Pitas, I. View-invariant action recognition based on artificial neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *23*, 412–424. [CrossRef] [PubMed]
15. Iosifidis, A.; Tefas, A.; Pitas, I. Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis. *Signal Proc.* **2013**, *93*, 1445–1457. [CrossRef]
16. Wang, L.; Qiao, Y.; Tang, X. Action recognition with trajectory-pooled deep-convolutional descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4305–4314.
17. Lei, J.; Li, G.; Zhang, J.; Guo, Q.; Tu, D. Continuous action segmentation and recognition using hybrid convolutional neural network-hidden Markov model model. *IET Comput. Vis.* **2016**, *10*, 537–544. [CrossRef]
18. Gkalelis, N.; Nikolaidis, N.; Pitas, I. View indepedent human movement recognition from multi-view video exploiting a circular invariant posture representation. In Proceedings of the IEEE International Conference on Multimedia and Expo 2009 (ICME 2009), New York, NY, USA, 28 June–3 July 2009; pp. 394–397.
19. Weinland, D.; Özuysal, M.; Fua, P. Making Action Recognition Robust to Occlusions and Viewpoint Changes. In *Computer Vision–ECCV 2010*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 635–648.
20. Souvenir, R.; Babbs, J. Learning the viewpoint manifold for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2008 (CVPR 2008), Anchorage, AK, USA, 23–28 June 2008; pp. 1–7.
21. Liu, J.; Shah, M. Learning human actions via information maximization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2008 (CVPR 2008), Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
22. Zheng, J.; Jiang, Z.; Chellappa, R. Cross-View Action Recognition via Transferable Dictionary Learning. *IEEE Trans. Image Proc.* **2016**, *25*, 2542–2556. [CrossRef] [PubMed]
23. Nie, W.; Liu, A.; Li, W.; Su, Y. Cross-View Action Recognition by Cross-domain Learning. *Image Vis. Comput.* **2016**. [CrossRef]

24. 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]
25. Chaaraoui, A.A.; Flórez-Revuelta, F. A Low-Dimensional Radial Silhouette-Based Feature for Fast Human Action Recognition Fusing Multiple Views. *Int. Sch. Res. Not.* **2014**, *2014*, 547069. [CrossRef] [PubMed]
26. Cheema, S.; Eweiwi, A.; Thurau, C.; Bauckhage, C. Action recognition by learning discriminative key poses. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 1302–1309.
27. Ahmad, M.; Lee, S.-W. HMM-based human action recognition using multiview image sequences. In Proceedings of the 18th International Conference on Pattern Recognition 2006 (ICPR 2006), Hong Kong, China, 20–24 August 2006; pp. 263–266.
28. Pehlivan, S.; Forsyth, D.A. Recognizing activities in multiple views with fusion of frame judgments. *Image Vis. Comput.* **2014**, *32*, 237–249. [CrossRef]
29. Chun, S.; Lee, C.-S. Human action recognition using histogram of motion intensity and direction from multiple views. *IET Comput. Vis.* **2016**, *10*, 250–257. [CrossRef]
30. Murtaza, F.; Yousaf, M.H.; Velastin, S. Multi-view Human Action Recognition using 2D Motion Templates based on MHIs and their HOG Description. *IET Comput. Vis.* **2016**. [CrossRef]
31. Hsieh, C.-H.; Huang, P.S.; Tang, M.-D. Human action recognition using silhouette histogram. In Proceedings of the Thirty-Fourth Australasian Computer Science Conference, Perth, Australia, 17–20 January 2011; Australian Computer Society, Inc.: Darlinghurst, Australia, 2011; pp. 213–223.
32. Rahman, S.A.; Cho, S.-Y.; Leung, M.K. Recognising human actions by analysing negative spaces. *IET Comput. Vis.* **2012**, *6*, 197–213. [CrossRef]
33. Sonka, M.; Hlavac, V.; Boyle, R. *Image Processing, Analysis, and Machine Vision*; Cengage Learning: Belmont, CA, USA, 2014.
34. Hu, M.-K. Visual pattern recognition by moment invariants. *IRE Trans. Inform. Theory* **1962**, *8*, 179–187.
35. Huang, Z.; Leng, J. Analysis of Hu's moment invariants on image scaling and rotation. In Proceedings of the 2010 2nd International Conference on Computer Engineering and Technology (ICCET), Chengdu, China, 6–19April 2010; pp. 476–480.
36. Jordan, A. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Adv. Neural Inform. Proc.g Syst.* **2002**, *14*, 841.
37. Qian, H.; Mao, Y.; Xiang, W.; Wang, Z. Recognition of human activities using SVM multi-class classifier. *Pattern Recognit. Lett.* **2010**, *31*, 100–111. [CrossRef]
38. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
39. Vapnik, V.N.; Vapnik, V. *Statistical Learning Theory*; Wiley: New York, NY, USA, 1998.
40. Kreßel, U.H.-G. Pairwise classification and support vector machines. In *Advances in Kernel Methods*; MIT Press: Cambridge, MA, USA, 1999.
41. Platt, J.C.; Cristianini, N.; Shawe-Taylor, J. Large Margin DAGs for Multiclass Classification. *Nips* **1999**, *12*, 547–553.
42. Dietterich, T.G.; Bakiri, G. Solving multiclass learning problems via error-correcting output codes. *J. Artif. Intell. Res.* **1995**, *2*, 263–286.
43. Cheong, S.; Oh, S.H.; Lee, S.-Y. Support vector machines with binary tree architecture for multi-class classification. *Neural Inform. Proc.-Lett. Rev.* **2004**, *2*, 47–51.
44. Hsu, C.-W.; Lin, C.-J. A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.* **2002**, *13*, 415–425. [PubMed]
45. Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2011**, *2*, 27. [CrossRef]
46. Manosha Chathuramali, K.; Rodrigo, R. Faster human activity recognition with SVM. In Proceedings of the 2012 International Conference on Advances in ICT for Emerging Regions (ICTER), Colombo, Sri Lanka, 12–15 December 2012; pp. 197–203.
47. Weinland, D.; Boyer, E.; Ronfard, R. Action recognition from arbitrary views using 3D exemplars. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–20 October 2007; pp. 1–7.

48. Reddy, K.K.; Liu, J.; Shah, M. Incremental action recognition using feature-tree. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 27 September–4 October 2009; pp. 1010–1017.
49. Lv, F.; Nevatia, R. Single view human action recognition using key pose matching and viterbi path searching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2007 (CVPR'07), Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
50. Vitaladevuni, S.N.; Kellokumpu, V.; Davis, L.S. Action recognition using ballistic dynamics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2008 (CVPR 2008), Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
51. Cherla, S.; Kulkarni, K.; Kale, A.; Ramasubramanian, V. Towards fast, view-invariant human action recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 2008 (CVPRW'08), Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
52. Wu, X.; Xu, D.; Duan, L.; Luo, J. Action recognition using context and appearance distribution features. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 489–496.
53. Burghouts, G.; Eendebak, P.; Bouma, H.; Ten Hove, J.M. Improved action recognition by combining multiple 2D views in the bag-of-words model. In Proceedings of the 2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Krakow, Poland, 27–30 August 2013; pp. 250–255.