Action Recognition Using Random Forest Prediction with Combined Pose-based and Motion-based Features

Ilktan Ar[1 2], Yusuf Sinan Akgul[2]

[1] Department of Computer Engineering, Kadir Has University, 34083, Cibali, Istanbul, Turkey
ilktana@khas.edu.tr
[2] Department of Computer Engineering, Gebze Institute of Technology, 41400, Gebze, Kocaeli, Turkey
akgul@bilmuh.gyte.edu.tr

Abstract

In this paper, we propose a novel human action recognition system that uses random forest prediction with statistically combined pose-based and motion-based features. Given a set of training and test image sequences (videos), we first adopt recent techniques that extract low-level features: motion and pose features. Motion-based features which represent motion patterns in the consecutive images, are formed by 3D Haar-like features. Pose-based features are obtained by the calculation of scale invariant contour-based features. Then using statistical methods, we combine these low-level features to a novel compact representation which describes the global motion and the global pose information in the whole image sequence. Finally, Random Forest classification is employed to recognize actions in the test sequences by using this novel representation. Our experimental results on KTH and Weizmann datasets have shown that the combination of pose-based and motion-based features increased the system recognition accuracy. The proposed system also achieved classification rates comparable to the state-of-the-art approaches.

1. Introduction

Human action recognition is an important topic in digital signal processing. The human action recognition systems analyze the image sequences or videos by using various approaches to predict the action type. This topic is very challenging because it does not only handle variations between scenes, people, and motion characteristics, but it also deals with performance issues based on low illumination, occlusions, perspective effects, etc. Robust and efficient recognition of actions is needed in various systems such as video surveillance, gesture recognition, virtual reality, behavior analysis, etc.

In a survey study, Gavrila [1] pointed out that the human action recognition will be one of the most popular research areas of computer vision. Approaches in this survey were generally based on 2D and 3D modeling of human body. In another survey, Wang *et al.* [2] summarized the studies based on human motion analysis between 1997-2001. They defined three issues in human motion analysis which were human detection, tracking, and activity understanding. Moeslund *et al.* [3] reviewed advances in human motion capture and analysis from 2000 to 2006. They mentioned that research in human tracking and pose estimation obtained successful results but much more work is needed and expected in the recognition of human actions.

Poppe [4] mentioned that human action recognition frameworks can be studied in two categories with respect to image representation and action classification techniques. Image representation techniques are divided into two categories: global and local. Local representations describe independent features obtained from local patches, interest points. On the other hand, global representations describe the feature characteristics for all the frames in the video.

The existing action recognition systems employ various low-level features such as pixel displacements, motion histogram images, optical flows, spatio-temporals, interest points, silhouettes, contours, edges, color, poses, texture, etc., and combinations of them. Ballan *et al.* [5] constructed descriptors based on the combination of optic flow vectors and 3D gradients on the interest points. They classified video sequences using these descriptors. Dhillon *et al.* [6] formed a motion representation which is based on the statistical values of optical flow vectors. Then a multi-class Adaboost classifier is fed by these representations and majority voting is applied to receive a classification decision for every sequence. Dondera *et al.* [7] developed a motion-based representation which can handle short-term and long-term actions efficiently. This representation models motion features as the probability distribution over the augmented Hough space. In order to classify the action type, the similarity among these representations is calculated. Baysal *et al.* [8] presented a method to recognize actions which use only pose information without any temporal information. Key poses which are a collection of line-pairs are extracted for every sequence and then they are employed to classify the given sequence with a majority voting algorithm. Motivated by the recent improvements on motion analysis and pose estimation, we argue that the combination of motion-based and pose-based features would be more beneficial than using them solely.

In this work, we show that an action recognition system which uses both motion and pose information in a compact representation can achieve better recognition rates. Therefore, we employ statistical methods to form a novel global representation from local (scale invariant contour-based) pose and local (3D Haar-like) motion features. Furthermore, we design a novel action recognition system which uses this compact representation to recognize the action type in the given image sequence. Note that our approach does not require any body-part segmentation, joint detection, interest point detection & tracking.

The remainder of the paper is organized as follow: Section 2 presents our action recognition system with further details in the consecutive subsections. Section 3 reports experiments and results. Finally, Section 4 concludes the paper.

2. The System

The proposed system takes videos which can be called image sequences as inputs and outputs the labels of the actions in these image sequences. Inside the proposed system, there are four different processes as shown in the Fig. 1. Local motion and pose information in the image sequences are extracted by adopting recent approaches. Then, this local information is combined into a novel global representation with statistical methods. Finally, the compact representations are fed into the Random Forest to obtain the action types in the given image sequence. Note that we assume that a video or an image sequence contains only one type of action performed by a person.

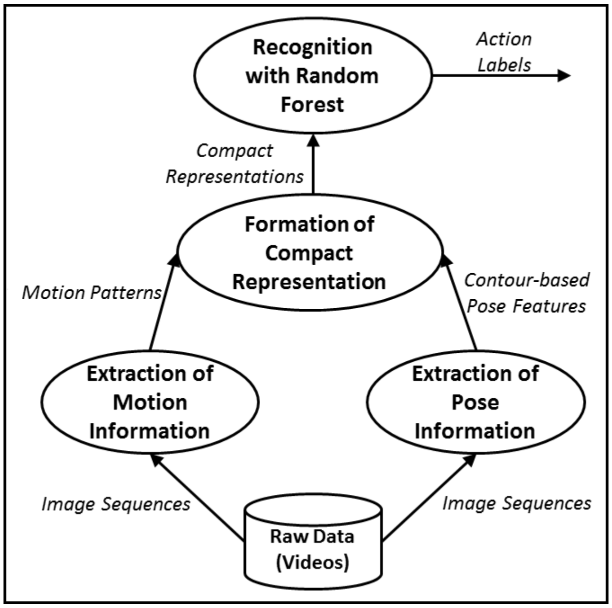
Recognition with Random Forest
Action Labels
Compact Representations
Formation of Compact Representation
Motion Patterns
Contour-based Pose Features
Extraction of Motion Information
Extraction of Pose Information
Image Sequences
Image Sequences
Raw Data (Videos)

Figure 1. The data flow diagram of the action recognition system.

2.1. Extraction of Motion Information

Motion information between consecutive frames define local motion features. We adopt our earlier method in [9] which uses 3D Haar-like wavelets to extract these local motion features.

Let IS denote an image sequence which contains images as $I_1...I_n$ where n defines the time order of the image in this sequence. We build cubic filters to calculate the local motion information between consecutive images in IS. Different cubic filters are needed to capture variations in the motion characteristics of actors. There are 24 different cubic filters available in our system. Examples of cubic filters are shown in Fig. 2. Note that changing the size of cubic filters in spatial and/or temporal domain forms different filters.

3D Haar-like features ($3DHFs$) in the consecutive images are the motion patterns which are dependent on spatial location, temporal location, and filter-type. $3DHFs$ are obtained with the convolution process ($*$) as

$$3DHF_t(x, y, f) = I_t(x, y) * CubicFilter_f, \qquad (1)$$

where t is the id (time order) of I and f is the id of the cubic filter from 1 to 24, x and y are the 2D coordinates of I. In this convolution process, all the pixel intensities in the black and the white region are summed separately and then the absolute difference between these sums is calculated as the output. Then, the local motion pattern (LMP) is formed by building the histograms of $3DHFs$ as

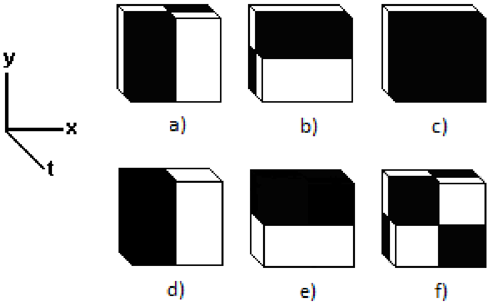

Figure 2. Examples of cubic filters which are used in the extraction of 3D Haar-like features.

$$LMP_t(f) = Histogram[3DHF_t(x, y, f)]. \qquad (2)$$

The $LMPs$ contain spatially independent motion patterns and represent the motion information between consecutive images where the consecutiveness is ranged by the filter type. The LMP is adequate to differentiate between short term actions (e.g. swinging of arm). The LMP can be defined as a vector with 256 bins for a given 8 bits gray-level images.

The motion information (MI) for a given image sequence IS is represented with a vector by merging $LMPs$ with concatenation processes ($||$) as

$$MI_t = LMP_t(f_1)||LMP_t(f_2)||...||LMP_t(f_{24}), \qquad (3)$$

where $f_1...f_{24}$ are the filter ids. The MI represents the local motion information between consecutive images using all the filter types. Note that, the motion information can be formed quickly by employing the integral volumes approach [10] for the calculation of 3D Haar-like features.

2.2. Extraction of Pose Information

A primitive action such as hand waving, running, bending, or etc. can be described as a sequence poses with pose descriptors. There are various descriptors available to represent human poses (line-pairs, histogram of oriented gradients, silhouettes, etc. [1, 3, 4]). Although pose extraction process requires preprocessing overhead, we believe that pose information is a valuable source for human action recognition tasks as in [11].

Cheema *et al.* [12] developed a human action recognition system based on pose representation from silhouettes. Although silhouette images for the image sequences are generally available for public human action recognition dataset, silhouettes can be obtained by simple background subtraction and threshold methods or complex background modeling and foreground prediction methods. Since we focus on pose representation, we assume that silhouette images are formed with the necessary preprocessing methods as in [12]. We adopt their pose descriptor which is formed by scale invariant contour features.

Pose representation starts with the calculation of center of mass (CM) in the binary silhouette image and continues with

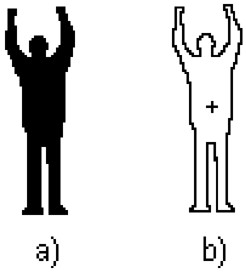

Figure 3. A silhouette image for an actor who is waving his two hands, is given in a) and the related image that demonstrates contour points, is given in b). '+' sign in b) indicates the center of mass for the silhouette in a).

the contour extraction. Next, these contour points (CPs) are ordered in a clockwise fashion starting with the contour which is on the horizontal-left of CM. A silhouette image for an actor who is waving his two hands, is shown in Figure 3a and the related image that demonstrates contour points and the CM (with a '+ ' sign) are shown in Figure 3b. Then, the Euclidean distances between CPs and CM are calculated and a distance vector (DV) is formed as

$$DV = [||CP_1 - CM||, ||CP_2 - CM||, ..., ||CP_s - CM||], \tag{4}$$

where s is the count of contour points. After that, each distance value in DV is scaled to a constant size c as $\widehat{DV}$ and $\widehat{DV}$ is normalized to $\overline{DV}$ as

$$\overline{DV[i]} = \frac{\widehat{DV[i]}}{\sum_1^s \widehat{DV[i]}}, \tag{5}$$

where i denotes the contour points and c controls the distance transform in terms of granularity of the features. Finally, we use these $\overline{DV}$ vectors and call them as pose information PI to represent pose-based features in our system.

2.3. Formation of Compact Representation

There are some situations where motion or pose information is not enough to recognize actions. We exemplify two different scenario to describe these shortcomings. In the first scenario, there are two different actions such as: waving right-hand and waving left-hand, where these actions are performed with the same actor. When we analyze the motion information, we observe similar representation for these actions such as there are horizontal displacements with the same magnitude for each hand. If we examine the pose information, we can distinguish the actions due to hand positions. In the last scenario, there are two different actions such as: walking and marching (paced walking). Although pose information about these actions are similar, motion information can distinguish them with displacement's magnitude and frequency. We believe that combining motion and pose information can tackle the similar shortcomings which are sampled in the previous scenarios.

Extraction of low-level features which represent local motion and pose information in consecutive and still images, is described in previous sections 2.1. and 2.2., respectively. Although these local representations are adequate to use in the recognition of different actions where the durations of the movements are too limited (4-8 frames) or the poses of the still images are very different, they are not adequate in the majority of actions. We need a more compact representation to understand the global motion and pose information about a given image sequence. For this task, we propose a novel global representation which is build by employing statistical methods.

Let an image sequence IS contain n images as $I_1...I_n$. Then, $MI_1...MI_n$ and $PI_1...PI_n$ are the related local pose and motion information, respectively. To build the compact global representation (CGR) for a given IS, we first calculate the global motion information (GMI) and the global pose information (GPI) with statistical methods.

GMI defines the dominant local motion patterns and the differences between them as

$$\mu(MI) = \frac{\sum_1^n MI_i}{n}, \tag{6}$$

$$\sigma(MI) = \frac{(MI_i - \mu(MI))^2}{n-1}, \tag{7}$$

$$GMI = \mu(MI)||\sigma(MI), \tag{8}$$

where μ shows the mean(average), σ shows the variance, and $||$ shows the concatenation process.

GPI defines the dominant pose of the actor and the variations between the remaining poses and the dominant pose. GPI is formed similar to GMI as

$$GPI = \mu(PI)||\sigma(PI). \tag{9}$$

CGR is the core representation of combination of motion and pose information in a holistic manner. This representation will be used in the further classification steps. The CGR is defined as a vector which is build by combining the global motion and the global pose information as

$$CGR = GMI||GPI. \tag{10}$$

2.4. Recognition with Random Forest

We need classifiers to recognize the human actions by the given compact representations. We prefer to employ the Random Forest classifiers for this task. The Random Forest can handle thousands of input variables and large dataset efficiently. Moreover, the Random Forest exhibits excellent performance and outperforms many other machine learning algorithms [13, 14].

Random Forest was introduced by Breiman [15] as a set of decision trees. Each decision tree in this forest behave like weak classifiers and come together to form a strong classifier. During training stages, nodes in the trees are split by randomized selection of features. This selection decreases the error rate in forest by decreasing the correlation among trees in the forest. Finally, each random tree in the forest grows and predicts the input test data's class label. The importance of variables are estimated at the end of training stage.

3. Experimental Results

We evaluate our system with two frequently used human action recognition databases: Weizmann dataset [16] and KTH dataset [17].

The Weizmann dataset [16] contains 90 low-resolution (180 x 144 pixels) video sequences of nine different people. Each person performs ten different actions such as bend, walk, run, wave one hand (wave1), wave two hand (wave2), jump, jumping jack (shortly jack), jump in place (pjump), skip, and gallop

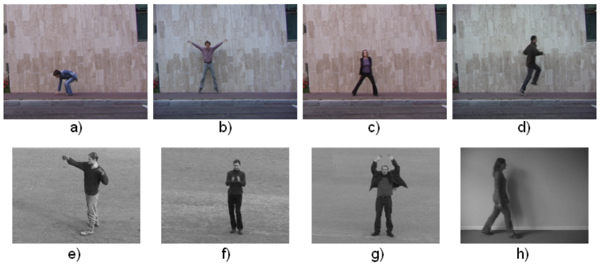

Figure 4. Sample frames of the Weizmann dataset (a-d) and the KTH dataset (e-h). Images are scaled with 0.4 factor.

sideways (shortly side). Sample images for the bend, jack, side, and skip are given in the Fig. 4. a, b, c, and d, respectively.

The KTH dataset [17] contains 599 low-resolution (160 x 120 pixels) video sequences of six different action classes such as walking, jogging, running, boxing, clapping, and waving. Each action is performed by 25 different actors in four different scenarios (indoor, outdoor, outdoor with different zoom levels, outdoors with different clothes). For each action classes there are 100 different video sequences except clapping (99 sequences).Sample images for the different scenarios and different actions are given in Fig. 4. where e) is a boxing image, f) is a hand clapping image, g) is a hand waving image, and h) is a walking image.

In the preprocessing stage, all the video sequences of the datasets are converted to image (frame) sequences. Next, RGB images are converted to 8-bits grayscale images. Then, noise reduction process in the images is accomplished with low-pass filtering techniques. Finally, morphological operators are used to form silhouette images.

In the experiments, we followed leave-one-actor-out approach to obtain the recognition accuracy of our system. Optimal parameters of Random Forest are estimated experimentally.

Table 1. Recognition results on the Weizmann dataset.

	wave1	wave2	bend	walk	run	jump	jack	pjump	side	skip
wave1	9	0	0	0	0	0	0	0	0	0
wave2	0	9	0	0	0	0	0	0	0	0
bend	0	0	9	0	0	0	0	0	0	0
walk	0	0	0	8	1	0	0	0	0	0
run	0	0	0	1	8	0	0	0	0	0
jump	0	0	0	0	0	8	0	0	0	1
jack	0	0	0	0	0	0	9	0	0	0
pjump	0	0	0	0	0	0	0	9	0	0
side	0	0	0	0	0	0	0	0	9	0
skip	0	0	0	0	1	1	0	0	0	7

We tested our system on Weizmann dataset to understand the benefits of combining motion and pose information for the human action recognition task. We used Eq.8 and Eq.9 to calculate separately, the representations of the global motion information (GMI) and the global pose information (GPI) for each image sequences. After that, we employed binary Support Vector Machines ($SVMs$) to obtain classification results with leave-one-actor-out approach. 79 of 90 image sequences were recognized successfully with an accuracy rate of 87.78% using only representations of GPI and 82 of 90 image sequences were recognized successfully with an accuracy rate of 91.11% using only representations of GMI, respectively. The most misclassified action pair was running and walking for the GPI based system and the most misclassified actions were running, jumping, and skipping for the GMI based system. Finally, we run our action recognition system with the $CGRs$ (Eq. 10) for each image sequence. Our system successfully recognized the actions with an overall accuracy rate of 94.44%. Although the misclassification errors were reduced by combining the motion and pose information effectively, there were some misclassification generally based on 'skip' action which is the combination of running and jumping actions. Further details of the results achieved by our system on the Weizmann dataset are given in Table 1.

Table 2. Recognition results on the KTH dataset.

	walk	jog	run	box	clap	wave
walk	95	3	2	0	0	0
jog	5	85	10	0	0	0
run	4	9	87	0	0	0
box	0	0	0	96	3	1
clap	0	0	0	2	96	1
wave	0	0	0	1	1	98

Table 3. Comparison of our system with the other systems in terms of action recognition accuracy.

System	**Dataset**	**Act.**	**Acc. (%)**
Bregonzio *et al.* 2012 [18]	Weizmann	10	96.66
Zhang & Tao 2012 [19]	Weizmann	10	93.87
Baysal *et al.* 2010 [8]	Weizmann	9	92.6
Ballan *et al.* 2009 [5]	Weizmann	10	92.41
Cheema *et al.* 2011 [12]	Weizmann	9	91.6
Dhillon *et al.* 2009 [6]	Weizmann	10	88.5
Thurau 2007 [20]	Weizmann	10	86.6
Bregonzio *et al.* 2012 [18]	KTH	6	94.33
Zhang & Tao 2012 [19]	KTH	6	93.5
Ballan *et al.* 2009 [5]	KTH	6	92.1
Baysal *et al.* 2010 [8]	KTH	6	91.5
Donderac *et al.* 2009 [7]	KTH	6	85.1
Dhillon *et al.* 2009 [6]	KTH	6	82.66
This paper (only motion)	Weizmann	10	91.11
This paper (only pose)	Weizmann	10	87.78
This paper (combined)	Weizmann	10	94.44
This paper (combined)	KTH	6	92.99

We also tested the performance of our system on the KTH dataset and obtained an overall accuracy of 92.99% for the recognition of six different action classes. The pose and motion information suffered by camera movements, image noise, textured background, illumination and scene variations, reduced the performance of our system on this dataset. The most misclassified actions were jogging and running which are very similar actions in terms of pose and motion information. Moreover, the misclassified actions can be grouped in two different sets such as walk-jog-run and box-clap-wave. This two groups can also be referred as upper-body and lower-body actions. We believe that addition of a body-part segmentation based approach to our system, can increase the overall recognition performance on this dataset. Further details of the recognition performance of our system on the KTH dataset are given with a confusion matrix in Table 2. Note that there is only 99 sequences available for clapping action.

We increase the recognition accuracy by using combined representation.The increments are between 3.3%-6.66% for the Weizmann dataset and between 3.84%-7.12% for the KTH dataset.Therefore, the experimental results on KTH and Weizmann datasets have shown that combination of pose and motion information is quite effective when the actions can not be classified by the motion or the pose information alone. Moreover, the comparison between different systems in the literature supports the hypothesis of feature fusion as in Table 3. Systems which are based on feature fusion [19, 18] like our system achieves better results than the systems which are based on a single type feature sets such as [6, 8, 12]. The proposed system achieved similar results within the state-of-the-art methods which are listed in Table 3. The small differences in terms of recognition accuracy on KTH dataset can be neglected due to the ceiling effect on this dataset [21].

4. Conclusions

In this work, we address the problem of human action recognition from image sequences. We propose a hypothesis which is based on the formation of a novel representation based on the global motion information and the global pose information can be advantageous for the action recognition tasks. For this hypothesis, we formed a novel representation by combining two recent local representations based on motion and pose information with statistical methods. We tested the efficiency of this representation by a novel system which employs Random Forests in the action recognition tasks.

Experimental results show that our system performed effectively in the action recognition task and achieved accuracy rates which can be comparable to the performance of state-of-the-art systems. Therefore, we found it beneficial to combine different source of information such as motion and pose knowledge instead of having a descriptor based on a single source of information.

5. References

[1] D.M. Gavrila, "The Visual Analysis of Human Movement: A Survey", *Computer Vision and Image Understanding*, vol. 73, no. 1, pp: 82-98, 1999.

[2] L. Wang and W.M. Hu and T.N. Tan, "Recent Developments in Human Motion Analysis", *Pattern Recognition*, vol. 36, no. 3, pp: 585-601, 2003.

[3] T.B. Moeslund and A. Hilton and V. Kruger, "A Survey of Advances in Vision-Based Human Motion Capture and Analysis", *Computer Vision and Image Understanding*, vol. 103, no. 3, pp: 90-126, 2006.

[4] R. Poppe, "A Survey on Vision-Based Human Action Recognition", *Image and Vision Computing*, vol. 28, no. 6, pp: 976-990, 2010.

[5] L. Ballan and M. Bertini and A. del Bimbo and L. Seidenari and G. Serra, "Recognizing Human Actions by Fusing Spatio-Temporal Appearance and Motion Descriptors", *Int. Conf. on Image Processing*, Cairo, Egypt, 2009, pp: 3569-3572.

[6] P.S. Dhillon and S. Nowozin and C.H. Lampert, "Combining Appearance and Motion for Human Action Classification in Videos", *Computer Vision and Pattern Recognition Workshop*, Miami, Florida, USA, 2009, pp: 22-29.

[7] Radu Donderac and David Doermann and Larry Davis, "Action Recognition Based on Human Movement Characteristics", *Int. Conf. on Motion and Video Computing*, Snowbird, Utah, USA, 2009, pp: 103-110.

[8] Sermetcan Baysal and Mehmet Can Kurt and Pinar Duygulu, "Recognizing Human Actions Using Key Poses", *The 20th Int. Conf. on Pattern Recognition*, Istanbul, Turkey, 2010, pp: 1727-1730.

[9] Ilktan Ar and Yusuf Sinan Akgul, "A Framework for Combined Recognition of Actions and Objects", *Int. Conf. on Computer Vision and Graphics*, Warsaw, Poland, 2012, pp: 264-271.

[10] Xinyi Cui and Yazhou Liu and Shiguang Shan and Xilin Chen and Wen Gao, "3D Haar-Like Features for Pedestrian Detection", *Int. Conf. on Multimedia and Expo*, Beijing, China, 2007, pp: 1263-1266.

[11] Angela Yao and Juergen Gall and Gabriele Fanelli and Luc Van Gool, "Does Human Action Recognition Benefit from Pose Estimation?", *The 22th British Machine Vision Conf.*, Dundee, Scotland, 2011, pp: 67.1-67.11.

[12] Shahzad Cheema and Abdalrahman Eweiwi and Christian Thurau and Christian Bauckhage, "Action Recognition by Learning Discriminative Key Poses", *Int. Conf. on Computer Vision Workshops*, Barcelona, Spain, 2011, pp: 1302-1309.

[13] Vladimir Svetnik and Andy Liaw and Christopher Tong and J. Christopher Culberson and Robert P. Sheridan and Bradley P. Feuston, "Random Forest: a Classification and Regression Tool for Compound Classification and QSAR Modeling", *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, 2003, pp: 1947-1958.

[14] David Meyera and Friedrich Leischa and Kurt Hornikb, "The Support Vector Machine Under Test", *Neurocomputing*, vol. 55, no. 1-2, 2003, pp: 169-186.

[15] Leo Breiman, "Random Forests", *Machine Learning*, 2001, pp: 5-32.

[16] Moshe Blank and Lena Gorelick and Eli Shechtman and Michal Irani and Ronen Basri, "Actions as Space-Time Shapes", *The 10th IEEE Int. Conf. on Computer Vision*, Beijing, China, 2005, pp: 1395-1402.

[17] Christian Schuldt and Ivan Laptev and Barbara Caputo, "Recognizing Human Actions: A Local SVM Approach", *17th Int. Conf. on Pattern Recognition*, Cambridge, UK, 2004, vol. 3, pp:32-36.

[18] Matteo Bregonzio and Tao Xiang and Shaogang Gong, "Fusing Appearance and Distribution Information of Interest Points for Action Recognition", *Pattern Recognition*, vol. 45, no.3, 2012, pp: 1220-1234.

[19] Zhang Zhang and Dacheng Tao, "Slow Feature Analysis for Human Action Recognition", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, 2012, pp: 436-450.

[20] Christian Thurau, "Behavior Histograms for Action Recognition and Human Detection", *The 2nd Conf. on Human Motion: Understanding, Modeling, Capture and Animation*, Rio de Janeiro, Brazil, 2007, pp: 299-312.

[21] Zan Gao and Ming-Yu Chen and Alexander G. Hauptmann and Anni Cai, "Comparing Evaluation Protocols on the KTH Dataset", *Human Behavior Understanding*, 2010, pp: 88-100.