# REAL-TIME ACTION RECOGNITION BASED ON CUMULATIVE MOTION SHAPES

*Marlon F. Alcântara, Thierry P. Moreira and Helio Pedrini*

Institute of Computing - University of Campinas
Campinas, SP, Brazil, 13083-852

## ABSTRACT

Although several methods for action recognition have been proposed in the literature, many of them have limitations in terms of applicability in real-life situations. Despite satisfactory accuracy rates achieved by a number of methods, an effective action recognition system requires workability in real time. However, this feature usually comes along with certain loss in accuracy. In this paper, we present a real-time action recognition method that achieves state-of-the-art accuracy. By accumulating shape information over a sliding window on the video frames, the method extracts and processes silhouettes with little computational effort. Simple descriptors are computed over the shapes and applied on a fast configuration of classifiers. Experiments are conducted on three public data sets and the results demonstrate the effectiveness of the method in terms of accuracy and speed.

***Index Terms***— Action Recognition, Motion Shape, Real Time

## 1. INTRODUCTION

Automated human action recognition is fundamental in surveillance tasks since human beings are susceptible to failure under stress and repetitive conditions. Moreover, recordings are often just stored, without any kind of verification, except in case of casualties.

An action consists of a single period of human movement – such as walking or taking steps, waving hands and collapsing – and can afterwards be used to infer the activities that occur in the video – for instance, person tracking, jumping a turnstile or convulsing [1]. The focus of this paper is on the domain of the actions.

One of the main challenges related to this context is the computational time required to process video frames, which often makes it impossible to apply action recognition in real life situations. In this work, we address this problem and show a solution that presents state-of-the-art accuracy and operates in real time.

A very popular approach to the action recognition problem employs Spatio-Temporal Interest Points (STIP) [2, 3, 4, 5, 6, 7, 8, 9], chosen as 3D corners or salient points. Some commonly used detectors are [10, 11, 12]. Bag-of-Visual-Words are usually constructed by clustering descriptors of the surroundings of the STIPs. Visual words individually encode only appearance information.

Various attempts have been proposed to improve classification rates by adding other types of data. In [4, 5], some additional geometric information is aggregated to the descriptors. In [8], a covariance matrix between the interest points is created. These methods have achieved impressive correct recognition rates. However, they usually present speed drawback, since the stages of finding interest points and computing their descriptors are time consuming.

The authors are grateful to FAPESP, CNPq and CAPES for the financial support.

Other popular approaches include shape analysis [1, 13, 14, 15, 16, 17, 18, 19]. This kind of method usually employs silhouette or pose for action recognition, often allowing it to be performed through a single static image. Most commonly, such works transform the shapes into signatures. 2D Radon transform is used in [17]. The centroid of the silhouette is used in [1, 14, 15, 18, 19] and a signature is generated by applying distance functions to each silhouette point in a radial scheme. Some advantages of shape analysis include its simplicity and easiness to compute representations, still being rich enough to represent actions. The drawback is that meaningful silhouettes may be difficult to acquire, relying mainly on background subtraction or frame difference – which usually fails when only a part of the body, or none, moves.

It is noticeable that the methods that work in real time have achieved smaller correct recognition rates than most of the others. Tables 1 and 2 show such fact by comparing these methods [1, 15, 18, 20, 21] to the others, which have only showed good results in the simplest data sets.

| Method | Data Set | |
|---|---|---|
| | KTH | Weizmann |
| Ryoo and Aggarwal (2009) [2] | 93.8 | - |
| Sun et al. (2009) [3] | 94.0 | 97.8 |
| Wang et al (2009) [20] | - | 93.3 |
| Ta et al. (2010) [4] | 93.0 | 94.5 |
| Raja et al. (2011) [13] | 86.6 | - |
| Hsieh et al. (2011) [14] | - | 98.3 |
| Bregonzio et al. (2012) [5] | 94.3 | 96.7 |
| Junejo and Aghbari (2012) [21] | - | 88.6 |
| Zhang and Tao (2012) [6] | 93.5 | 93.9 |
| Onofri, Soda (2012) [7] | 97.0 | - |
| Chaaraoui et al. (2013) [15] | - | 90.3 |
| Ji et al. (2013) [22] | 90.2 | - |
| Guo et al. (2013) [8] | 98.5 | 100 |
| Moghaddam and Piccardi (2013) [23] | - | 96.8 |
| Alcântara et al. (2013) [1] | - | 94.6 |

**Table 1**. Comparison of correct prediction rates (in percentage) for KTH and Weizmann data sets.

In this paper, we contribute with a lightweight real-time descriptor for human action recognition based on motion shapes. It has achieved state-of-the-art accuracy in three popular public data sets (Weizmann [25], KTH [26] and MuHAVi [16]) comparable to slower, more complex descriptors.

The segmentation process and the input information processing is explained in Section 2. The proposed action recognition methodology is described in Section 3. Section 4 presents and discusses the results obtained by applying the method to three public data sets.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

| Method | Data Set | | |
|---|---|---|---|
| | MuHAVi | MuHAVi8 | MuHAVi14 |
| Wu et al. (2010) [9] | 69.2† | - | - |
| Moghaddam and Piccardi (2010) [24] | 80.4 | - | - |
| Singh et al. (2010) [16] | - | 82.4 | 97.9 |
| Karthikeyan et al. (2011) [17] | 88.2 | - | - |
| Cheema et al. (2011) [18] | - | 95.6 | 86.0 |
| Moghaddam and Piccardi (2013) [23] | 92.0 | - | - |
| Chaaraoui et al. (2013) [15] | - | 97.1 | 91.2 |
| Chaaraoui and Flórez-Revuelta (2013) [19] | - | 100 | 98.5 |

†Experiments conducted by [17].

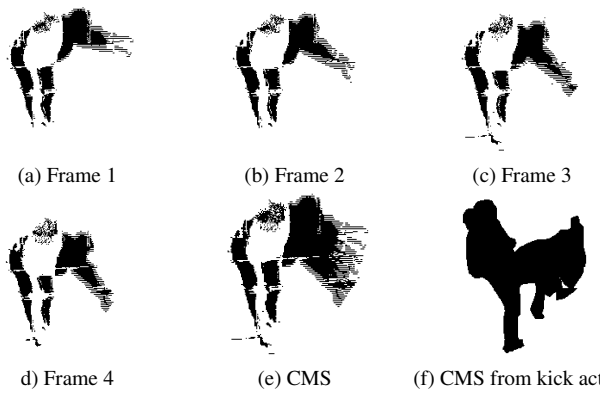
**Table 2**. Comparison of correct prediction rates (in percentage) for MuHAVi and its manually annotated sub-datasets, MuHAVi14 and MuHAVi8.

Section 5 concludes the paper and includes some directions for future works.

## 2. CUMULATIVE MOTION SHAPES

A motion shape is the moving part of an action in a given frame of a video sequence. In a perfect scenario, the motion shapes correspond to the silhouettes of the actors. However, extracting good silhouettes from videos is a challenge and there is still no algorithm that extracts them well enough in an acceptable time. We refer to them as motion shapes since errors are acceptable. For example, if a person moves only their arms, the only movement segmented would be the shape of an arm.

For the foreground segmentation, the background subtraction described in [27] is used. Frames with outlier values are discarded; a frame is considered outlier when the bounding box shows little movement, no movement at all, or when the movement occupies a very large portion of the frame – usually due to camera movement or a sudden change of light conditions; an adaptive background subtraction is capable of relearning the background. To avoid losing important parts of the shape, a morphological reconstruction is done, so that disconnected parts of the actor are put together. It can be seen in Figures 1 (a)-(d) that parts from the back, leg and feet are reattached to the shape.



(a) Frame 1 (b) Frame 2 (c) Frame 3

d) Frame 4 (e) CMS (f) CMS from kick action

**Fig. 1**. (a)-(d) Examples of poorly extracted foreground from smash object action; (e) CMS from joining previous images; (f) CMS from kick action. All images extracted from MuHAVi data set.

The cumulative motion shapes (CMS) are the simple union of all the motion shapes extracted from the video frames in a given sliding window. The CMS for the $k$-th frame of a sequence ($M_k$) is given by Equation 1, where $n$ is the size of the sliding window and $S_i$ is the motion shape of the $i$-th frame.

$$M_k = \bigcup_{i=k-n}^{k} S_i \tag{1}$$

Different actions often have common poses. The CMS adds temporal information to poses without raising dimensionality, therefore, neither requiring more processing nor memory usage. In the case of mis-segmentation, the CMS sometimes gather the broken portions of movement, setting up a meaningful shape. Examples of CMS are shown in Figures 1 (e)-(f).

## 3. METHODOLOGY

The methodology proposed in this work is illustrated in Figure 2. Its main stages are explained in details as follows.

Initially, Cumulative Motion Shapes (CMS) are extracted (step 'a') from the video stream. A person, when is performing some action, can move a body part more or less than other person performing the same action, so the entire cloud of points can give false clues about the action. Because of that, there must be an adequate sampling over this cloud (step 'b') and, additionally, it is needed a way to correlate the points between distinct frames.

### 3.1. Bounding-box key points and interest point strategy

To acquire the interest points (step 'c'), the main idea is to select extreme points on CMS; the extreme points are the nearest points from some key points fixed on the bounding box. The key points are equally distributed along the bounding box sides, as shown in Figure 3(a). The number key points can be parameterized, however, it is the same over all video streams used in the training process.

The four corners of the bounding box are denoted as $c_a$, $c_b$, $c_c$ and $c_d$. The $k$-th subdivision ($p_k$) between two adjacent corners, $c_x$ and $c_y$, is represented by Equation 2, where $D$ is the number of bounding box subdivisions on the edge between the two corners. The points are treated as vectors for the operations.

$$p_k = \frac{k.(c_x - c_y)}{D} + c_x \tag{2}$$

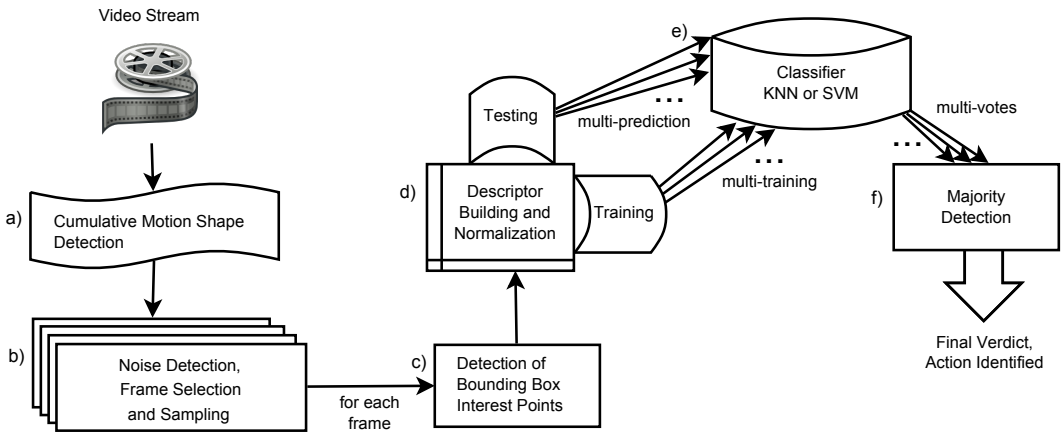


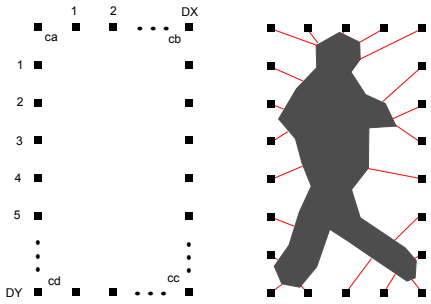

**Fig. 2**. Diagram illustrating the main stages of the proposed methodology.

The bounding box sides are divided by a fixed number, which does not need to be the same for all four sides. Eventually, it can be interesting to use a distinct number for horizontal and vertical sides. This is because the CMS can have more information disposed in vertical than the horizontal direction.

A video sequence contains one CMS for each frame. The CMSs from the sequences are sampled for the training. A constant number of valid CMSs is considered to create the final descriptor (step 'd'). There is no precedence order between distinct CMS from the same video stream.

Extracting multiple samples from a same sequence helps learning actions starting from any part of its period – for example, a walking action may start with the two feet together or with after a step has already been taken.

The centroid of the CMS is found and the final descriptor is constructed from the set of coordinates of the interest points previously computed by using the centroid as the origin and normalized in relation to the bounding box.




(a) bounding box subdivision (b) feature extraction

**Fig. 3**. Interest points. (a) scheme illustrating how the control points are found in the bounding box; (b) characterization of the interest points by the distances from the CMS to the control points.

### 3.2. Multi-training

A set of CMSs are extracted from each video sequence. Each one is used in the training independently. In other words, each video stream will populate the same classification machine (step 'e') $N$ times, where $N$ is a constant value defined by the user.

Similarly, when a prediction is made, a set of descriptors sampled from the test sequence is computed. Each one is used to a distinct prediction. Then, the classifier answers with which action each descriptor fits well. Each prediction works as a vote and the one that appears most of the times (step 'f') is the final verdict. The creation of the descriptor is the same both for training and testing.

## 4. EXPERIMENTAL RESULTS

Experiments were conducted on three data sets: Weizmann, KTH and MuHAVi. All the experiments were performed by using the proposed multi-training in K-Nearest Neighbor (K-NN) and multi-class Support Vector Machines (SVM). The latter one, in some cases, achieved better accuracy, however, K-NN works considerably faster.

Classification time through K-NN is more sensitive to the number of training samples, whereas, for the (multi-class) SVM, the classification time is more sensitive to the number of trained classes. Hence, the number of training samples extracted from each sequence may vary depending on the classifier.

All the measured times were found by taking the average of 5 runs. The machine used was an i7 (3.5 GHz) computer and no parallelism mechanism was implemented. The feature extraction was coded in C++ programming language with *OpenCV* library. The classification code was written separately in *R* package through the machine learning libraries *e1071* and *kernlab*.

### 4.1. Weizmann Data Set

Weizmann [25] is an action data set consisting of 10 classes, with 9 actors performing each action, sometimes with some actors performing them more than once, totalling 93 videos. The frames are captured at 25 FPS, size of $180 \times 144$ pixels. All the actions occur in the same, static, background.

The Weizmann data set has a total of 5,701 frames, 228.04 seconds at 25 FPS. The extraction of the features from the frames took a total time of 4.85 seconds – average of 1,175.95 FPS. The CMS are constructed by using 4 frames and the number of samples extracted for each sequence for K-NN is 30, while the best value is 23 for SVM. Accuracy rates and classification times are shown in Table 3.

| | SVM | K-NN |
|---|---|---|
| Classification time (ms) | 3 | 3 |
| Accuracy rate (%) | **96.77** | 95.70 |

**Table 3**. Accuracy rates (in percentage) and classification time (in milliseconds) for Weizmann data set.

### 4.2. KTH Data Set

KTH [26] is an action data set consisting of 6 classes, with 25 actors performing each action, in 4 different scenes, with the exception of one person, that perform one action (*hand clapping*) in only 3 scenes, totalling 599 videos. The frames were captured at 25 FPS, size of $160 \times 120$ pixels. Most videos have camera movement – zooming, panning and tilting. Camera movements are serious threats for descriptors based on silhouettes or motion shapes whatsoever, making this data set a challenge for the method.

The KTH data set has a total of 289,715 frames, 11375.32 seconds at 25 FPS. The extraction of the features from the frames took a total time of 1347.38 seconds – average of 215.02 FPS. The CMS are constructed using 12 frames and the number of samples extracted for each sequence is 40. Accuracy rates and classification times are shown in Table 4.

| | SVM | K-NN |
|---|---|---|
| Classification time (ms) | 43 | 23 |
| Accuracy rate (%) | **90.11** | 88.78 |

**Table 4**. Accuracy rates (in percentage) and classification time (in milliseconds) for KTH data set.

### 4.3. MuHAVi Data Set

MuHAVi [16] (Multicamera Human Action Video Data) is a multiview data set consisting of 17 classes, with 7 actors performing each action, totalling 119 videos. The actions occur in a closed scenario, with 8 cameras surrounding it. Since this work does not focus on multiviewing, only one camera is used, the camera 4, which captures the action from the side. The frames were captured at 25 FPS, size of $720 \times 576$ pixels.

The optimal number of shapes accumulated to build the CMS is 40. It is higher than the other data sets since the actions in MuHAVi are more complex and need much more computational time to take place. Therefore, the actions require many frames to characterize them. For instance, the *climb ladder* action consists of a person walking towards a ladder, climbing it up, down, and then walking away from it. The number of samples used per sequence used with K-NN is 40, whereas 50 were used with SVM. Accuracy rates are shown in Table 5.

The MuHAVi has a total of 134,085 frames, 5,368.16 seconds at 25 FPS. The extraction of the features from the frames demanded a total time of 2,850.29 seconds – average of 47.04 FPS. Classification times are shown in Table 5.

The data set has a subset of manually annotated sequences (MuHAVi-MAS), in which the frames are binary images of the silhouette locations. It is divided into 14 primitive actions and it is usually called MuHAVi14 in the literature. This subset, however, has some actions that vary only in direction – for example, run left and run right – that are rearranged together, forming another subset with 8 classes, called MuHAVi8. The actions were annotated from cameras 3 and 4; in this case, again, only camera 4 is used.

The manually annotated subsets consist of shorter sequences, with simpler actions. Therefore, their actions are better described by CMS constructed by fewer frames. Moreover, since the segmentation was done manually, the CMS loses their restoration importance. The number of frames used in the CMS is 5 and the number of samples used per sequence is 12. Accuracy rates are shown in Table 5.

Since MuHAVi8 and MuHAVi14 have the same video sequences, in different classes, their feature extraction computational time results are the same. The entire sub-datasets have 3,969 frames, 158.76 seconds at 25 FPS. The total extraction of the features from the frames demanded an average time of 19.54 seconds – average of 203.12 FPS. The classification times differ depending on the machine, since the data sets have different numbers of classes. Computational times for classification are shown in Table 5.

| | Accuracy (%) | | Time (ms) | |
|---|---|---|---|---|
| Data Set | SVM | K-NN | SVM | K-NN |
| MuHAVi | 84.87 | **89.08** | 259 | 6 |
| MuHAVi14 | 91.18 | **94.12** | 48 | 1 |
| MuHAVi8 | 98.58 | **100** | 14 | 1 |

**Table 5**. Accuracy rates (in percentage) and classification time (in milliseconds) for MuHAVi data set.

## 5. CONCLUSIONS

Once individual motion silhouettes can not provide sufficient information to learn video actions, the method introduced and discussed in this paper is based on cumulative motion shapes (CMS), which aggregate temporal information to silhouettes.

Actions can be performed by different actors and with distinct starts. Because of that, correlating an order to CMS can be a complex task. Multi-training avoids such problem by using each CMS independently in the training process and considering each one such as a unique vote in the prediction judgment. The multi-training and judgment process are responsible for providing higher accuracy to the system.

The method was tested on three commonly used data sets and produced state-of-the-art accuracy (Tables 1 and 2), even on KTH data set (Table 4), which contains unrealistic camera movement, especially for surveillance, where cameras are mostly fixed. Experiments also demonstrated proper results on MuHAVi (Table 5), which is a difficult data set, since some sequences contain people walking around in the background or arranging the scenario and most videos include parts when the actor performs actions different from the one labeled (mostly walking), increasing inter-class similarities along the sequences. Finally, the method obtained high accuracy rates for Weizmann data set (Table 3) with impressive speed rate.

The real-time execution achieved by the method is due to the fast acquisition for CMS over a window of frames and the efficient interest points selection by using the bounding-box nearest points. Hence, the developed method can be applied in behavior analysis and surveillance.

Directions for future work include the evaluation of our description scheme in other classification problems, once it can produce results as good as STIP appearance-based descriptor used in many works, such as [2, 3, 4, 5, 6, 7, 8, 9].

## 6. REFERENCES


[1] M. F. de Alcântara, T. P. Moreira, and H. Pedrini, "Motion Silhouette-Based Real Time Action Recognition," in *18th Iberoamerican Congress on Pattern Recognition*, Havana, Cuba, Nov. 2013, vol. 8259, Lecture Notes in Computer Science, pp. 471–478.

[2] M. S. Ryoo and J. K. Aggarwal, "Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities," in *International Conference on Computer Vision*, Kyoto, Japan, 2009, pp. 1593–1600.

[3] X. Sun, M. Chen, and A. Hauptmann, "Action Recognition via Local Descriptors and Holistic Features," in *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 58–65.

[4] A.-P. Ta, C. Wolf, G. Lavoue, A. Baskurt, and J. M. Jolion, "Pairwise Features for Human Action Recognition," in *International Conference on Pattern Recognition*, Istanbul, Turkey, 2010, pp. 3224–3227.

[5] M. Bregonzio, T. Xiang, and S. Gong, "Fusing Appearance and Distribution Information of Interest Points for Action Recognition," *Pattern Recognition*, vol. 45, no. 3, pp. 1220–1234, Mar. 2012.

[6] Z. Zhang and D. Tao, "Slow Feature Analysis for Human Action Recognition," *Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 436–450, 2012.

[7] L. Onofri and P. Soda, "Combining Video Subsequences for Human Action Recognition," in *International Conference on Pattern Recognition*, Tsukuba, Japan, 2012, pp. 597–600.

[8] K. Guo, P. Ishwar, and J. Konrad, "Action Recognition From Video Using Feature Covariance Matrices," *Image Processing*, vol. 22, no. 6, pp. 2479–2494, 2013.

[9] C. Wu, A. H. Khalili, and H. Aghajan, "Multiview Activity Recognition in Smart Homes with Spatio-Temporal Features," in *International Conference on Distributed Smart Cameras*, Atlanta, Georgia, 2010, pp. 142–149.

[10] I. Laptev, "On Space-Time Interest Points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[11] A. Oikonomopoulos, I. Patras, and M. Pantic, "An Implicit Spatiotemporal Shape Model for Human Activity Localization and Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 27–33.

[12] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features," in *International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005, pp. 65–72.

[13] K. Raja, I. Laptev, P. Perez, and L. Oisel, "Joint Pose Estimation and Action Recognition in Image Graphs," in *International Conference on Image Processing*, Brussels, Belgium, Sept. 2011, pp. 25–28.

[14] C. H. Hsieh, P. Huang, and M. D. Tang, "The Recognition of Human Action Using Silhouette Histogram," in *Australasian Computer Science Conference*, Mark Reynolds, Ed., Perth, Australia, 2011, vol. 113, pp. 11–16.

[15] A. Chaaraoui, P. Climent-Prez, and F. Fórez-Revuelta, "Silhouette-based Human Action Recognition using Sequences of Key Poses," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799 – 1807, 2013, Smart Approaches for Human Action Recognition.

[16] S. Singh, S. A. Velastin, and H. Ragheb, "MuHAVi: A Multicamera Human Action Video Dataset for the Evaluation of Action Recognition Methods," in *Advanced Video and Signal Based Surveillance*, 2010, pp. 48–55.

[17] S. Karthikeyan, U. Gaur, B. S. Manjunath, and S. Grafton, "Probabilistic Subspace-based Learning of Shape Dynamics Modes for Multi-view Action Recognition," in *International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 1282–1286.

[18] S. Cheema, A. Eweiwi, C. Thurau, and C. Bauckhage, "Action Recognition by Learning Discriminative Key Poses," in *International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 1302–1309.

[19] A. A. Chaaraoui and F. Flórez-Revuelta, "Human Action Recognition Optimization based on Evolutionary Feature Subset Selection," in *Genetic and Evolutionary Computation Conference*, New York, NY, USA, 2013, pp. 1229–1236.

[20] S. Wang, K. Huang, and T. Tan, "A Compact Optical Flow-based Motion Representation for Real-Time Action Recognition in Surveillance Scenes," in *International Conference on Image Processing*, Cairo, Egypt, Nov. 2009, pp. 1121–1124.

[21] I. N. Junejo and Z. A. Aghbari, "Using SAX Representation for Human Action Recognition," *Journal of Visual Communication and Image Representation*, vol. 23, no. 6, pp. 853–861, Aug. 2012.

[22] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[23] Z. Moghaddam and M. Piccardi, "Training Initialization of Hidden Markov Models in Human Action Recognition," *Automation Science and Engineering*, vol. 36, no. 99, pp. 1–15, 2013.

[24] Z. Moghaddam and M. Piccardi, "Histogram-Based Training Initialisation of Hidden Markov Models for Human Action Recognition," in *International Conference on Advanced Video and Signal Based Surveillance*, Boston, MA, USA, 2010, pp. 256–261.

[25] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," in *International Conference on Computer Vision*, Beijing, China, 2005, pp. 1395–1402.

[26] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," in *17th International Conference on Pattern Recognition*, Cambridge, UK, 2004, vol. 3, pp. 32–36 Vol.3.

[27] P. Kaewtrakulpong and R. Bowden, "An Improved Adaptive Background Mixture Model for Real-Time Tracking with Shadow Detection," in *European Workshop on Advanced Video Based Surveillance Systems*, London, UK, 2001, vol. 5308.