3D Spatial-Temporal View based Motion Tracing in Human Action Recognition

Silambarasi R, Suraj Prakash Sahoo, Samit Ari

***Abstract*—The paper presents an extended approach of Motion History Image to trace the human motions in a video for recognizing the human actions. The video is represented as a 3D volume space and the trace of human motions are projected onto the three different views called 3D spatio-temporal plane. The extended view traces both the human shape and movement in different directions over the time. The Histogram of Oriented Gradients (HOG) features are extracted over all the projection plane which gives more distinct features for the action classification. Since HOG features data has high dimensionality, the optimal feature subset is selected by using the feature selection techniques. Finally, the Support Vector Machine (SVM) of multi-class classifier is used to identify the various actions of human. The various experiments are conducted on benchmark dataset KTH and results shows that the proposed method improves the action recognition rate compared to the existing methods.**

***Index Terms*—Human Action Recognition, 3D Spatio-Temporal Plane, Motion History Image, HOG, PCA, Feature Selection, SVM Classification.**

I. Introduction

Human action recognition is a well developing research area in computer vision. It plays a significant role in automatic intelligence based video surveillance, detection of abnormal activities, military, content based video retrieval, human-computer interaction, and robot learning and control etc. The recognition of various real life activities of human being is a challenging task with the presence of complex background structure, partial occlusion, low level illuminations, camera motions, and complex human actions. The human action comprises of several sub-actions which describes the motion on several parts of the body. The different activities may poses similar sub-actions which increases the complexity of action recognition.

The recognition of human action from the video is a generalized process of action classification from the extracted features. Action representations are important because they don't depend on human tracking [1] for action detections. In general, the feature can be classified as local features and global features. The local features are extracted by determining the interest points. The spatio-temporal local features [2]–[6] represent the video as a number of independent patches. The patches are the rectangular blocks extracted around the interest points where the eloquent changes are detected. The common interest point detector, Harris and Forstner interest points detector [4], detects the noticeable changes in the spatial location in terms of intensity level as well as the direction. Whereas, the global features describes the Region of Interest (ROI) of an image, which is derived from edges, silhouettes or optical flow. The common strategy of global features extraction is a Grid-based representation, enhances the efficiency of feature description even though the video has the presence of occlusion and noises. Histogram of Oriented Gradients (HOG) [7] and Histogram of Optical Flow (HOF) [8] are mostly used descriptors for extracting grid-based global features.The extracted local features are modelled by using the Codebook [9] and Bag-of-Words (BoW) [2], [3], [10]–[13] methods. Liu *et al.* [2] proposed a framework for the multiple view human action recognition with extended BoW representation. Another method of describing high level motions, the Motion History Image (MHI) [5], [14] based interest regions are detected in which the video is represented as a 3D volume space [3]. Samanta *et al.* represents video as 3D space time facet model and detects the STIP. Tian *et al.* [5] used the Motion History Image (MHI) as action representation.

The local features may lead to loss of significant interest points when the foreground and background don't have much variation in the intensity level. The local features are less sensitive to occlusion and noise but it is required to have adequate number of interest points to avoid loss of information. These local features contains information about the low level motions and high level motions are not detected. In order to detect the high level motions of human, the global features are used. However, the HOG feature extraction techniques gives large number of features which lead to increase of classification error due to the presence of redundant and irrelevant information. The proposed method overcomes above issues by calculating the Region of Interest by edge detection and higher order temporal derivatives. And the effective HOG features are extracted from the 3D spatio-temporal planes which extend the Motion History Image concept into three different planes. Then the Principal Component Analysis (PCA) followed by feature selection techniques are applied to reduce the irrelevant features and it increases the classification accuracy.

The paper is organized as, Section II describes about the detail of the proposed approach. Section III discusses the experimental results and comparison of the proposed approach against the early reported results on action recognition. Finally, section IV concluded the discussion.

Silambarasi R is with Department of Electronics and Communication Engineering, National Institute of Technology Rourkela, (email: simbu2494@gmail.com).

Suraj Prakash sahoo is with Department of Electronics and Communication Engineering, National Institute of Technology Rourkela (email: surajprakashsahoo@gmail.com).

Dr. Samit Ari is the assistant professor in Department of Electronics and Communication Engineering, National Institute of Technology Rourkela (email: samit.ari@gmail.com).

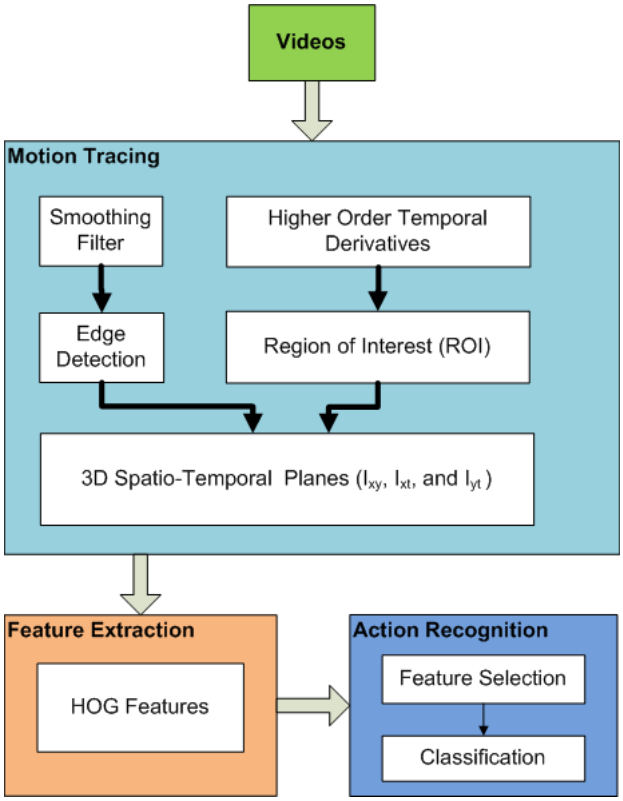

Fig. 1. Flowchart of Proposed Human Action Recognition

II. Proposed Framework

The proposed work is divided into three sections and the workflow is shown in Fig. 1. The detailed explanation has been given in the subsequent subsections. In this Proposed method, ROI is calculated from the edge detection and higher order temporal derivatives. The features are determined around the interest point and the feature selection techniques are used to improve the classification accuracy.

In traditional method, the Motion History Image was created and traces are projected onto a single plane (X-Y plane) which contains all the motion information over the time T. The proposed work extends the concept of Motion History Image (MHI) in 3D spatio-temporal view. The extended MHI projection contains the trace information of human motion in 3D volume space (shown in Fig. 2) with the three different planes. It increases the distinct features between the different actions and improves the learning efficiency.

A. Motion Tracing

In general, the video contains both background and foreground motion information. The detected background motions are due to clutters, slight illumination changes from frame to frame where as the foreground motions are due to object actions. Such background motions are considered as a noisy motions which reduces the accuracy of human action recognition. In order to remove the noisy motions the Smoothing

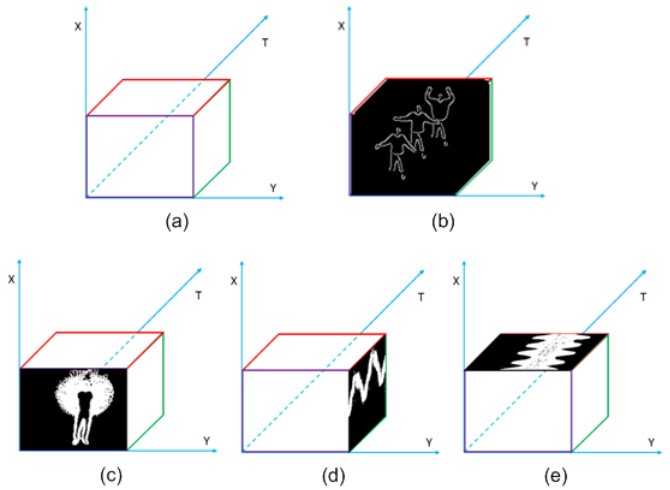

Fig. 2. 3D Volume representation of a video and the extended motion history image. The figure (a) shows the axis formation of 3D planes (b) modeled sequence of frames in 3D volume space (c) projection on XY plane (d) projection on XT plane (e) projection on YT plane

Filter, more specifically a Median Filter with window size 3×3, is applied on all the frames of the video data. The human shape is detected by using the *canny* edge detection [15] mechanism over the smoothed frames. Then the Motion History Image is generated by the process of projection of human shape from all the frames onto a single plane (XY plane).

1) Region of Interest: To improve the recognition rate, the features are extracted over the interest regions. The interest regions are defined by the significant changes in the intensity level over the frames which can be determined by higher order temporal derivatives. The detected frames are binarized by applying the threshold mechanism with an empirical hard threshold value which eliminates the noisy motions. The interest regions are described by the pixel value of 1 in the binarized frames whereas the remaining regions have been assigned a value of 0. These binarized frames in 3D space are then projected onto the spatio-temporal planes XT and YT.

2) 3D Spatio-Temporal Planes: The video is represented as three dimensional volume in spatio-temporal space with coordinates X (height), Y (width) and T (time), shown in Fig. 2(a). Instead of representing the original video, the combination of edge detected frames and higher order temporal derivatives of the frames are represented in 3D space, shown in Fig. 2(b). This 3D volume space is projected onto the planes of three different views XY, XT and YT, shown in Fig. 2(c,d,e). The view of this 3D spatio-temporal planes extend the single view of Motion History Image into three different view. The example of 3D spatio-temporal planes of various actions from KTH dataset are shown in Fig. 3. It clearly shows that traces are different for different actions. Eventhough the shape of the motions are similar for different actions, varying stride length distinguishes the various actions.

Thus, the idea of extending the MHI in multiple views based on 3D spatio-temporal planes distinguishes motion of different

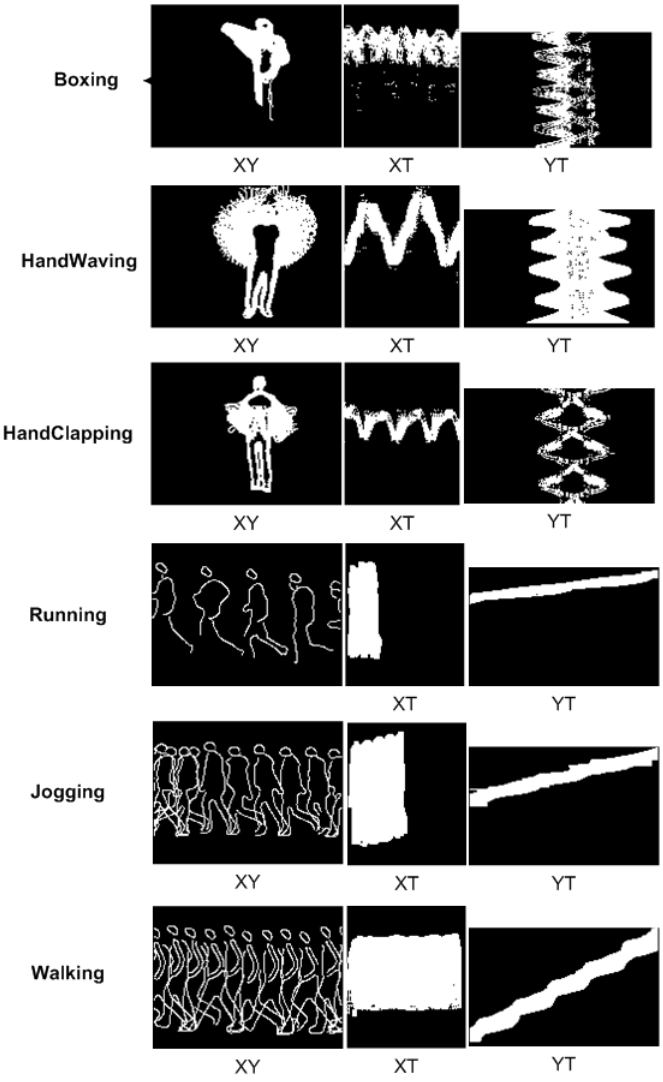

Fig. 3. 3D SpatioTemporal Planes XY, XT and YT of different actions of KTH dataset

human actions effectively. It enables the robustness of feature extraction for classification with minimized error rate.

B. HOG Feature Extraction

The Histogram of Oriented Gradients (HOG) features are computed by the orientation of edge intensity gradients. The HOG features represents the shape of an object as well as the direction information of the edge.

The sobel filter is used to calculate the gradients $dx(x,y)$ and $dy(x,y)$ in x and y direction. By using this directional gradients, the magnitude $M(x,y)$ and orientation $\theta(x,y)$ can be defined as

$$M(x,y) = \sqrt{dx(x,y)^2 + dy(x,y)^2} \tag{1}$$

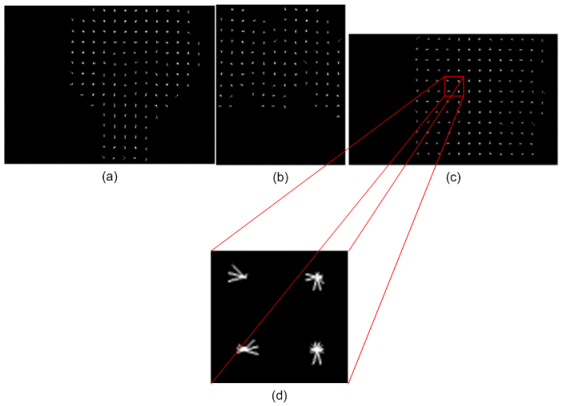

Fig. 4. HOG features description for action HandWaving from KTH dataset (a) Plane XY (b) Plane XT (c) Plane YT (d) Detail of a portion of Plane YT

$$\theta(x,y) = \begin{cases} \tan^{-1}(\frac{dy(x,y)}{dx(x,y)}) - \pi & \text{when dx(x,y)<0 and dy(x,y)} < 0 \\ \tan^{-1}(\frac{dy(x,y)}{dx(x,y)}) + \pi & \text{when dx(x,y)<0 and dy(x,y)} > 0 \\ \tan^{-1}(\frac{dy(x,y)}{dx(x,y)}) & \text{elsewhere} \end{cases} \tag{2}$$

In the proposed work, the HOG features are extracted over all the 3D spatio-temporal planes XY, XT and YT as shown in Fig. 4. The extraction of HOG features over the 3D planes are more effective because the plane contains only the interest regions such as the edges where the significant changes are occurred.

The HOG feature length N is calculated as

$$N = Blocks\ Per\ Image \times Block\ Size \times Number\ Of\ Bins \tag{3}$$

$$Blocks\ Per\ Image = \left\{ \frac{\frac{Image\ Size}{Cell\ Size} - Block\ Size}{Block\ Size - Block\ Overlap} + 1 \right\} \tag{4}$$

where, *Cell Size* is $[8\ 8]$ denotes the size of HOG cell, *Block Size* is $[2\ 2]$ denotes the number of cells in block, *Block Overlap* is half of Block Size denotes the number of overlapping cell between adjacent blocks and *Number Of Bins* is 12 denotes the number of directions. Finally, the features extracted over all the planes are combined together.

C. Action Recognition

The various human actions are recognized by training a Machine Learning Model from the extracted features and assigning the class label to the corresponding action class. The HOG features extracted from the 3D spatio-temporal planes are having a length more than ten thousands in number. These features may contains redundancy and irrelevant information, and it may reduces the classification accuracy. The combination of dimension reduction and feature selection techniques are employed which can increase the recognition rate.

1) Feature Selection: The PCA is a well known technique for the dimensionality reduction. It performs an orthogonal linear transformation of the data into a new low dimensional subspace which is linear and has the greater variances. Kobayashi *et al.* [7] applied PCA techniques for reducing the dimension of the HOG features. The another technique of dimension reduction is the feature selection in which the suboptimal feature subset is selected instead of the linear transform.This feature selection techniques are applied over the PCA transformed data. In this paper, the filter based feature selection techniques such as Pearson Correlation and Spearman correlation are applied. The classification results of PCA and feature selection techniques with various feature length are shown in Table I.

2) Classification: The multi-class SVM is most widely used classifier which gives greater classification accuracy. The SVM can support a high dimensional data by creating the maximal hyperplane which separates the non-overlapping classes. Since, the human action recognition with HOG features has high dimensional data, the multi-class SVM classifier is employed [3], [11], [16].

III. Experimental Results and Discussion

The experiments are conducted on benchmark dataset KTH in the environment, Matlab 2015b, 2.10 GHz Intel(R) Pentium(R) CPU B950 and Memory 2GB. It consists of 600 videos with six classes of human action such as walking, running, jogging, hand waving, hand clapping, and boxing. In each class there are 100 videos taken by 25 persons in four directions in various scenarios (outdoor with different cloths, different zooming level, indoor, camera motion, background motion). Each video has a size of 160×120 in spatial resolution and the number of frames taken for the experiment is 100. The projected 3D spatio-temporal planes XY, XT and YT has the size of 120×160, 120×100 and 100×160 respectively.

The HOG features extracted over the planes XY, XT and YT have feature length of 12768, 7392 and 10032 respectively, calculated by using the equation. 3 and 4. In total, 30192 features are extracted for each video sample. The extraction of 30192 features may contain redundancy and irrelevant features which reduces the classification accuracy. The PCA based dimensionality reduction techniques are implemented to reduce the feature dimensions and it reduces the error rate of classification. The various experiments are conducted for PCA with feature length 10 to 600 and the classification accuracy is shown in Fig. 5.

The multiclass SVM is used as the classifier with the kernel of Radial Basis Function. The reduced dimensionality of feature length 250 shown better classification results. Further the sub-optimal feature subsets are selected by using the various feature selection techniques such as Pearson Correlation and Spearman Correlation. The experiments of feature selection are performed with various feature length 10 to 250, with increment 10, over the features transformed by the PCA of length 550 and the classification accuracy is shown in Fig. 6.

Table I shows the classification accuracy of various feature reduction techniques. From Table I, it is clear that, feature

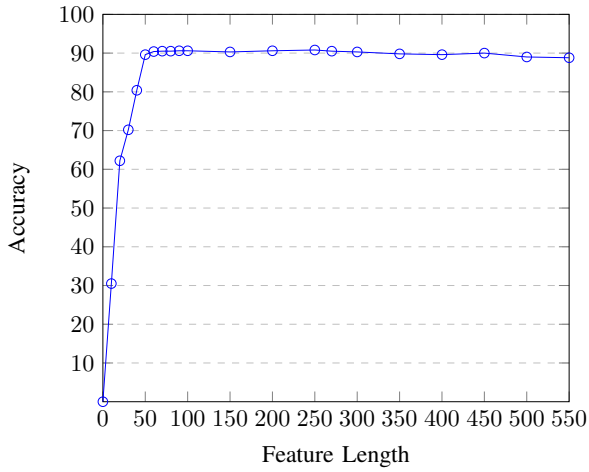

Fig. 5. Classification accuracy of PCA with various feature length

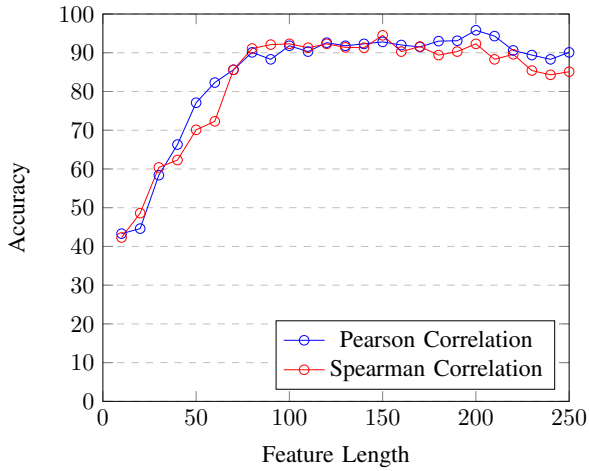

Fig. 6. Classification Accuracy of Feature Selection Techniques

selection techniques when added to PCA, performs better but, at the same time effective feature length is also increasing. Feature length of 200 can be compromised in comparison to 100 as it leads to very less computational complexity. Therfore, PCA with feature selection technique is preferable over simple PCA. PCA+Pearson correlation is giving better accuracy in comparison to PCA+Spearman correlation, so the former one is used for the feature selection procedure.

TABLE I. Classification accuracy of feature reduction techniques

Technique	Feature Length	Accuracy
PCA	250	90.6%
PCA+Pearson Correlation	200	95.8%
PCA+Spearman Correlation	150	94.5%

The proposed method gives better classification accuracy of 95.8%.and the corresponding confusion matrix for the classification result over the features subset selected by the Pearson Correlation of length 200 is shown in Table II. The

TABLE II. CONFUSION MATRIX FOR KTH DATASET

Action	Boxing	Hand clapping	Hand waving	Jogging	Running	Walking
Boxing	**0.99**	0.01	-	-	-	-
Hand clapping	-	**0.98**	0.02	-	-	-
Hand waving	0.01	0.01	**0.98**	-	-	-
Jogging	-	-	-	**0.91**	0.07	0.02
Boxing	-	-	-	0.07	**0.93**	-
Boxing	-	-	-	0.04	-	**0.96**

Table III shows the comparisons of classification accuracy with the existing methods of Human Action Recognition. It shows that the proposed method improved the recognition rate compared to the existing method.

TABLE III. COMPARISON WITH EXISTING METHODS

Method	Accuracy
Rodriguez et al. [17]	88.6%
Zhang et al. [16]	95.3%
Sadek et al. [18]	93.5%
Modarres et al. [19]	94.8%
Li et al. [14]	93.5%
Zhen et al. [12]	94.1%
Samanta et al. [3]	94.9%
Proposed Method	**95.8%**

IV. CONCLUSION

This paper presents an efficient approach i.e. 3D Spatial-Temporal View based Motion Tracing for recognition of various human actions like walking, running, jogging, hand waving, clapping, and boxing. The 3D spatio-temporal planes have extended the MHI by analyzing the shape and movement of human actions over time in multiple views. Thus the proposed method gives more distinct features between different classes of human actions with robust. The feature selection techniques have been added to traditional feature reduction technique such as PCA, to extract optimal feature subset. A comparison study of feature selection techniques have been analyzed. The effect of feature length on the overall accuracy have been studied. The experimental results show that the proposed method performs better compared to earlier reported techniques. The proposed method provides an accuracy of 95.8% which is better than state-of-art methods.

In future, the study can be extended to other feature descriptors such as HOF, Scale Invariant Feature Transform (SIFT), so that the method can handle more complex videos. For the same purpose the region of interest selection procedure can also be enhanced to work in more complex environment.

ACKNOWLEDGMENT

The work is supported by Media Lab Asia (Visvesvaraya Ph.D. Scheme for Electronics and IT under the department of DeitY government of India. The work is done at Department of Electronics and Communication Engineering, NIT Rourkela, India.

REFERENCES

[1] S. P. Sahoo and S. Ari, "Automated human tracking using advanced mean shift algorithm," in *Communications and Signal Processing (ICCSP), 2015 International Conference on*. IEEE, 2015, pp. 0789–0793.

[2] A.-A. Liu, Y.-T. Su, P.-P. Jia, Z. Gao, T. Hao, and Z.-X. Yang, "Multiple/single-view human action recognition via part-induced multitask structural learning," *IEEE transactions on cybernetics*, vol. 45, no. 6, pp. 1194–1208, 2015.

[3] S. Samanta and B. Chanda, "Space-time facet model for human activity classification," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1525–1535, 2014.

[4] I. Laptev, "On space-time interest points," *International journal of computer vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[5] Y. Tian, L. Cao, Z. Liu, and Z. Zhang, "Hierarchical filtered motion for action recognition in crowded videos," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 3, pp. 313–323, 2012.

[6] G. K. Yadav, P. Shukla, and A. Sethfi, "Action recognition using interest points capturing differential motion information," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1881–1885.

[7] T. Kobayashi, A. Hidaka, and T. Kurita, "Selection of histograms of oriented gradients features for pedestrian detection," in *International conference on neural information processing*. Springer, 2007, pp. 598–607.

[8] J. Perš, V. Sulić, M. Kristan, M. Perše, K. Polanec, and S. Kovačič, "Histograms of optical flow for efficient representation of body motion," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1369–1376, 2010.

[9] W. Zhou, C. Wang, B. Xiao, and Z. Zhang, "Human action recognition using weighted pooling," *IET Computer Vision*, vol. 8, no. 6, pp. 579–587, 2014.

[10] M. R. Amer and S. Todorovic, "Sum product networks for activity recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 4, pp. 800–813, 2016.

[11] M. M. Moussa, E. Hamayed, M. B. Fayek, and H. A. El Nemr, "An enhanced method for human action recognition," *Journal of advanced research*, vol. 6, no. 2, pp. 163–169, 2015.

[12] X. Zhen and L. Shao, "A performance evaluation on action recognition with local features," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 4495–4500.

[13] F. Moayedi, S. Dashti, R. Boostani, and Z. Azimifar, "Learning sparse shape bases for human action recognition," in *Electrical Engineering (ICEE), 2015 23rd Iranian Conference on*. IEEE, 2015, pp. 755–760.

[14] C. Li, Y. Liu, J. Wang, and H. Wang, "Combining localized oriented rectangles and motion history image for human action recognition," in *Computational Intelligence and Design (ISCID), 2014 Seventh International Symposium on*, vol. 2. IEEE, 2014, pp. 53–56.

[15] K. Xu, X. Jiang, and T. Sun, "Human activity recognition based on pose points selection," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2930–2834.

[16] T. Zhang, J. Liu, S. Liu, C. Xu, and H. Lu, "Boosted exemplar learning for action recognition and annotation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 853–866, 2011.

[17] M. Rodriguez, "Spatio-temporal maximum average correlation height templates in action recognition and video summarization," 2010.

[18] S. Sadek, A. Al-Hamadi, G. Krell, and B. Michaelis, "Affine-invariant feature extraction for activity recognition," *ISRN machine vision*, vol. 2013, 2013.

[19] A. F. A. Modarres and M. Soryani, "Body posture graph: a new graph-based posture descriptor for human behaviour recognition," *IET Computer vision*, vol. 7, no. 6, pp. 488–499, 2013.