

Human Action Recognition using Ensemble of Shape, Texture and Motion features

¹Aditi S. Jahagirdar, ²Manoj S. Nagmode

¹MIT College of Engineering, Pune

aditi.jah@gmail.com

²Government College of Engineering and Research, Avsari, Pune

Manoj.nagmode@gmail.com

Abstract

Even though many approaches have been proposed for Human Action Recognition, challenges like illumination variation, occlusion, camera view and background clutter keep this topic open for further research. Devising a robust descriptor for representing an action to give good classification accuracy is a demanding task. In this work, a new feature descriptor is introduced which is named 'Spatio Temporal Shape-Texture-Motion' (STSTM) descriptor. STSTM feature descriptor uses hybrid approach by combining local and global features. Salient points are extracted using Spatio Temporal Interest Points (STIP) algorithm which are further encoded using Discrete Wavelet Transform (DWT). DWT coefficients thus extracted represent local motion information of the object. Shape and texture features are extracted using Histogram of Oriented Gradient (HOG) and Local Binary Pattern (LBP) algorithms respectively. To achieve dimensionality reduction, Principal Component analysis is applied separately to three types of features. Feed Forward Neural Network is employed to perform the classification. The proposed algorithm is extensively tested using well known KTH, Weizmann and UCF sports datasets. The performance of proposed method is found to be better than many methods mentioned in the literature.

Keywords: Action recognition, DWT, HOG, Motion Feature, Shape Feature, STSTM, Texture Feature.

1.Introduction

In last decade, rapid growth in cheaper and easily available video recording equipment has generated huge video data. To develop automatic video analysis system has become challenging task for researchers working in Computer Vision. Automatic Human Action Recognition is important part of Video Analytics and finds application in wide area of intelligent video surveillance Systems, Human machine interface, Robotics, Interactive Video games, Healthcare system etc. Actions vary from simple gestures like lifting a hand, to very complicated activity like fighting, which is series of simple gestures. Action recognition becomes more complex for interaction between two humans or interaction between human and object. Crowd behavior detection is another type of Action Recognition.

Action classification becomes more demanding as every individual can perform many actions and every human being can perform same action in different ways. This increases intra class variation. Features extracted from simple actions like walking and jogging have very less variation, increasing inter class similarity [1]. Complexity of action recognition goes on increasing as number of human beings and objects involved in that action goes on increasing [3].

A basic HAR system involves preprocessing of frames, segmentation to find Region of Interest (ROI), feature extraction and classification. Every step of this process can be achieved in multiple ways making it possible to have various approaches for HAR. To improve the classification accuracy, multiple features, describing different properties of a video are extracted and different feature fusion techniques are explored [2]. In recent literature, Deep learning approach is implemented for achieving good accuracy and to avoid difficulty of feature selection, feature extraction and feature fusion [4].

In this work we focus on various features and fusion of these features to improve accuracy of Action Classification. Three basic types of features used for classification are shape based, motion based and texture based

Rest of the paper is organized as follows: Section (2) gives Related work. Section (3) explains basic background theory about features used. Section (4) gives proposed algorithm in detail. Section (5) elaborates experimental setup. Section (6) discusses results and section (7) gives Conclusion of the work.

2.Related Work

In recent literature, a variety of methods have been proposed by researchers for solving the problem of Action Recognition. Most of the methods proposed can be divided in two parts: detecting the Salient points or Region of Interest (ROI) and then describing the ROI with feature.

Maryam N., Al-Berry and others [5, 6] have proposed use of 3D stationary wavelet transform for obtaining local and global features for action recognition. Wavelet coefficients represent multi scale and directional information of motion pattern. In [7], Al-Berry and others proposed use of Local Binary Pattern (LBP) features along with SWT to get better accuracy of classification.

Aryanfar, Alihossein, et al [8] proposed use of DWT for dimensionality reduction. Distance feature is calculated for each frame and then DWT is applied on matrix of all frames thus generated to find approximate coefficients. These coefficients are further used for classification.

In [9], Thanikachalam, V., and K. K. Thyagarajan proposed a method using Accumulate motion Image (AMI) and Motion History Image (MHI). DWT coefficients extracted from MHI are used as spatial features along with LBP extracted from AMI. Hu moments are used further for forming a feature vector. Muhammad Hameed Siddiqi et al. proposed use of DWT coefficients for feature extraction. After applying 4 level DWT, stepwise linear discriminant analysis is used to find key features from huge number of feature.

Siddiqi, Muhammad Hameed, et al [10], proposed use of 4 level Discrete Wavelet Transform to extract the features. Stepwise Linear Discriminant Analysis is used to reduce dimensionality of features. In [11] shape, texture and motion features are used together with satisfactory results. Multimodal features which include audio and video are used for recognition in [12] with good results. In paper [13], a novel local image descriptor is introduced, which extracts the Histograms of Second Order Gradients which describe the curvature related geometric properties of object. In this work our focus is on action classification in low resolution videos.

3.Basic background Theory

The proposed system evaluates the performance of shape, texture and motion features used together. Spatio Temporal Interest points (STIP) are extracted before finding the motion features. For extracting STIPs, parameters proposed in [16] are used. Histogram of Oriented Gradients (HOG) is used to find shape information of the action. HOG is implemented using parameters given by Dalal & Triggs in [17]. Texture feature is extracted using Linear Binary Pattern (LBP) algorithm proposed in [18]. To find motion features, Discrete Wavelet Transform (DWT) is used. Basic concept of these features is given here.

3.1 Spatio temporal Interest Points

Spatio Temporal Interest Point detector proposed by Dollar et al. is used to find salient points. It is similar to Harris 3D detector proposed by Laptev but uses Gaussian filter in spacial domain and Gabor filter in temporal domain. As Gabor filter is linear separable filter, it is sensitive to periodic changes in pixel intensities. Thus it can detect significant motion of object.

Spatial regions which exhibit high variation in image intensity are termed as corners in 2d image. Same concept is extended in 3D to take into account temporal behavior of signal to find STIP. STIPs are detected at various spatial and temporal scales. Figure 1(a) shows detected STIPs on sample frame from KTH dataset.

3.2 Histogram of Oriented Gradient

Histogram of Oriented Gradients (HOG) is a feature descriptor introduced by Dala and Triggs in CPVR conference in 2005 [17] for person detection. The image is divided into small connected parts called cells. Equation 1 gives formula used to find magnitude and angle of the gradient.

$$|G_r| = \sqrt{P_x^2 + P_y^2} ; \varphi = \arctan \frac{P_y}{P_x} \quad (1)$$

where G_r represents magnitude and φ represents angle of the gradient. P_x and P_y represent X and Y derivative of image P computed by applying discrete derivative mask. The histogram of intensity gradient of pixels in this cell is computed considering bins of angles. These histograms are concatenated to derive the feature vector. Further normalization is applied considering bigger parts images called blocks. This makes HOG robust to illumination changes. Figure 1(b) shows HOG features extracted from sample frame of KTH dataset. Till date HOG is most favorite feature descriptor considered in local feature approaches used for human detection.

3.3 Local Binary Pattern

Binary Pattern (LBP) descriptor is an efficient way to describe texture of an image. In LBP, each pixel is compared with neighborhood pixels and labeled using a binary pattern generated by converting negative values to 0 and positive values to 1. Equation 2 gives formula used to find LBP of each pixel. In this work circular neighborhood of 8 pixels is used as in [14].

$$LBP_c = \sum_{x=0}^x S(I_x - I_c) * 2^x ; S(p) = \begin{cases} 1, & \text{if } p \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where I_x is gray value of x th pixel in circular neighborhood and I_c denotes the gray value of central pixel. Thus each pixel is represented using stream of eight bits. After LBP labeled image is obtained, LBP histogram is derived. Figure 1(c) shows LBP features of sample frame from KTH dataset.

3.4 Discrete Wavelet Transform

In this work Discrete Wavelet Transform (DWT) is used to compute the motion feature. DWT is known to detect edges of an object which helps in tracking the object in videos. Subset of wavelet coefficients is used to define the movement of the object. In this work Daubechies wavelet is used to find wavelet coefficients. Equation 3 gives formula to find coefficients using wavelet transform.

$$Coeff = A_3 + [(D_h)_i + (D_v)_i + (D_d)_i] \quad (3)$$

where A_3 gives 3rd level average coefficients and D_h , D_v and D_d give horizontal, vertical and diagonal coefficients respectively. Figure 1(d) shows first level DWT features of sample frame from Weizmann dataset

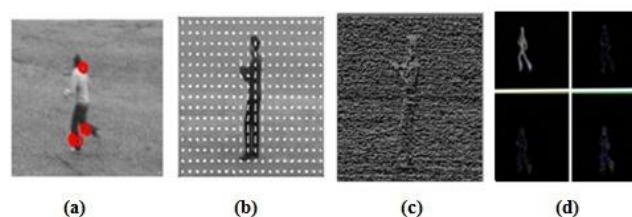


Figure 1: (a) STIP features (b) HOG features (c) LBP Features (d) DWT Coefficients

4. Proposed System

In this work, STSTM feature descriptor is proposed which uses combination of Shape, Texture and Motion features. HOG, LBP and DWT algorithms are implemented to extract these features respectively. BoF approach is used to combine these three types of features. STSTM feature descriptor is a hybrid descriptor which syndicates local and global feature concept.

4.1 Feature Extraction

Here motion feature is extracted around the STIPs which represent local features whereas HOG and LBP are extracted at frame level and represent global features. Test video is first read and converted to frames. Each frame is treated as an image and then features are extracted from each frame. STIP is used to localize motion in every frame using Cuboids method. Each feature is represented in a matrix form where columns correspond to feature values and rows correspond to number of frames. Detail algorithm is given in Figure 2.

4.2 Feature selection

Size of the features which are extracted using HOG, LBP and DWT is in thousands. Many times huge size of feature vector as compared to training samples, hampers the classification accuracy. Therefore selection of number of features is a challenging task. Various dimensionality reduction techniques like LDA, PCA, ICA, ISOMAP etc. have been proposed in literature for selecting good features.

In this work, Principal Component Analysis (PCA) is used for feature selection. PCA finds eigenvalues and eigenvectors and arranges them in descending order starting with first principal component. Thus features having maximum information are placed at the starting of the matrix. Finding number of principal components required to get good classification accuracy is still a challenging task and has to be found empirically.

4.3 Feature descriptor

After feature selection is done, the three features viz. HOG, LBP and DWT are in matrix form. To characterize these features, first order moment of each feature matrix is calculated. First order moment defines center of mass. Thus every matrix is converted to a row vector. Row vector of HOG, LBP and DWT thus obtained are concatenated to form a feature descriptor.

Feature descriptors of training data are stored and given as input to classifier along with feature descriptor of Test data. Labeling of training data is done manually.

Input: T_{video} subject to trained data

Output: Sel_{class}

Begin

 Read T_{video} ;

$F[] \leftarrow convert_{frames}()$;

for $k = 1$ to F_{max}

$_mST_n \leftarrow SPT_{feature}(F_k)$;

end for

for $k = 1$ to F_{max}

$_kHM_n \leftarrow Motion_{feature}(_mST_n)$;

$_kPHM_n \leftarrow filter_{pca}(_kHM_n)$;

end for

for $k = 1$ to F_{max}

$_kHS_n \leftarrow Shape_{feature}(F_k)$;

$_kPHS_n \leftarrow filter_{pca}(_kHS_n)$;

end for

for $k = 1$ to F_{max}

$_kHT_n \leftarrow Texture_{feature}(F_k)$;

$_kPHT_n \leftarrow filter_{pca}(_kHT_n)$;

end for

$f_{vector} \leftarrow featurecat(_1PHM_n, _1PHS_n, _1PHT_n)$;

$Sel_{class} \leftarrow Classifier(f_{vector})$

End

Figure 2: Steps of Proposed Algorithm in detail

Input to the system is a test video. It is assumed that training features extracted from the training videos are stored in a file and labeling of training videos is done. Test video T_{video} is first read and converted to frames. SPTfeature function extracts STIPs from all the frames and stores in the form of cuboids. Motionfeature function finds DWT coefficients and stores in matrix form in kHM_n ,

where k represent number of frames. PCA is applied to reduce the dimensionality and retain the features with highest information in $kPHMn$. Same process is applied to obtain $kPHSn$ using Shapefeature function. HOG algorithm is used to extract shape feature. $kPHTn$ is obtained by applying Texturefeature function which uses LBP algorithm followed by PCA. Mean of $kPHMn$, $kPHSn$, and $kPHTn$ is found and is represented as a row array $1PHMn$, $1PHSn$, and $1PHTn$. These three feature vectors are concatenated to form a final feature descriptor f_{vector} . Feed Forward Neural Network (FFNN) classifier is used to classify the Test video using the f_{vector} obtained.

5. Experimental Setup

5.1 Dataset Used

The algorithm is tested on three well-known action datasets, Weizmann, KTH and UCF sports. Weizmann dataset have 10 simple actions like walking, running, bending, jumping in one place, skipping etc., performed by 9 different actors and recorded in controlled environment. KTH dataset has 6 actions viz. walking, running, jogging, boxing, clapping and hand waving performed by 25 different actors, recorded in 4 different scenarios. UCF sports dataset includes total of 150 sequences having 10 different sport actions. The collection represents a natural pool of actions featured in a wide range of scenes and viewpoints. This dataset is a benchmark dataset for action recognition and used by many researchers. Figure 3 shows sample frames of datasets used.

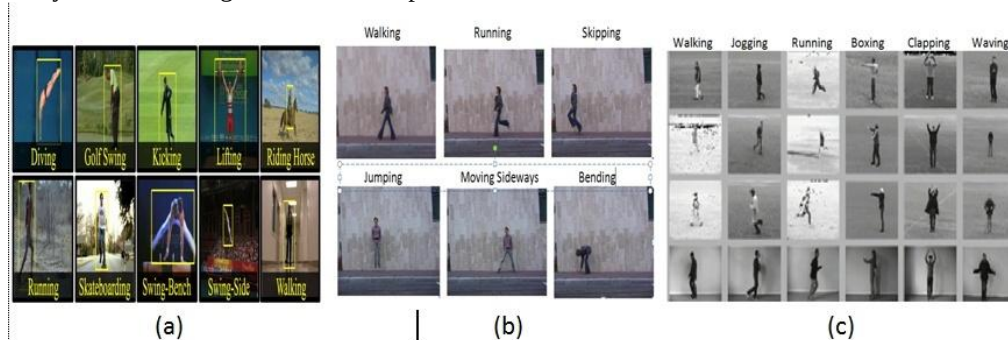


Figure 3: Sample frames (a) UCF Sports action dataset (b) Weizmann dataset (c) KTH dataset

5.2 Experimental Results

For testing the proposed algorithm, Recognition Accuracy is considered as a performance parameter. Experimentation is done in two setups. In first scenario, Weizmann and KTH datasets are used for Training and Testing in various combinations as given here:

1. Weizmann dataset is used for training as well as testing (Combination 1)
2. KTH dataset is used for training as well as testing (Combination 2)
3. Weizmann dataset is used for training and KTH dataset is used for testing (Combination 3)
4. Both the datasets are used for training as well as testing (Combination 4)

For this evaluation, fusion of Shape, Texture and Motion feature is used. For combination 3, only the actions common in KTH and Weizmann datasets are used. The results showed that high accuracy is achieved if samples from same dataset are used for training and testing. 92.8% accuracy is obtained when samples from one dataset are used for training and other for testing. It highlights the fact that, even if there is variation in back ground, illumination and view angle of video, proposed system gives comparable accuracy. Figure 4 shows graph of recognition accuracies achieved with combination of KTH and Weizmann datasets used for training and testing.

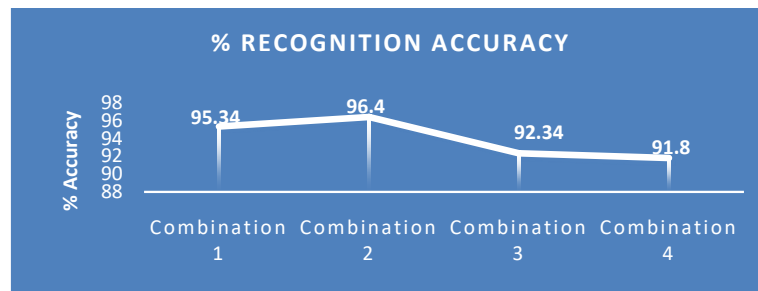


Figure 4: Recognition Accuracy achieved with Combination of datasets used for Training and Testing

In second setup, effect of combination of shape, texture and motion feature is studied. Shape-Texture, Shape-Motion, Texture-Motion and Shape-Motion –Texture feature combinations are used separately to find Recognition accuracy. Feed Forward Neural Network is trained using 15 hidden layers. For finding the Recognition accuracy, 80% samples are used for training and 20% samples are used for testing. Figure 5 shows graph of comparison of recognition accuracy achieved with combination of features.

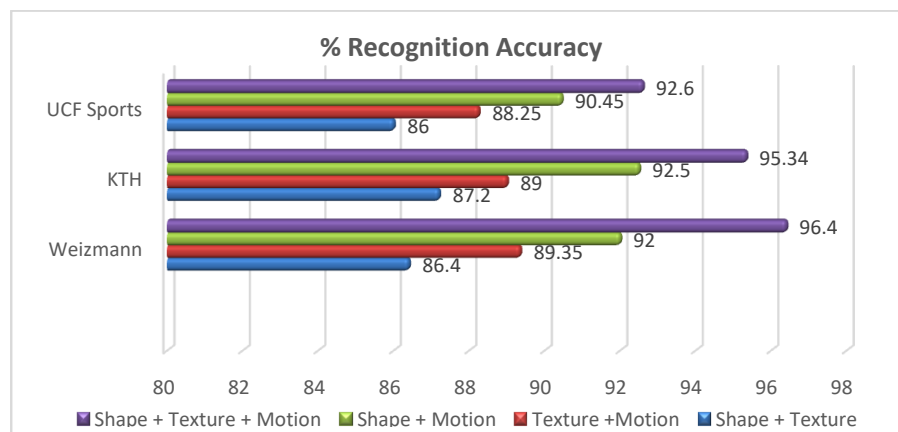


Figure 5: Comparison of Recognition Accuracy achieved with Combinations of Features

It is observed that Shape, Texture and Motion features used together give maximum classification accuracy. As the data sets used in this work are specially recorded for action recognition in controlled environment, variation in texture is less. So the texture feature does not contribute much in classification as compared to shape and motion feature. Comparable accuracy is achieved all datasets used for testing. For further validating the usefulness of the proposed approach, results obtained on three datasets are compared with results obtained by other state of the art methods. Recognition accuracy is considered as performance measure. It is observed that for all three datasets used for testing, comparable accuracy is achieved by proposed approach. Table 1 shows comparison of results attained with proposed method with state of art methods available in literature.

6. Conclusion

This work presents a new algorithm, STSTM, for action classification. It is a hybrid method which integrates local and global feature approaches. Local features give salient points of frame and global features describe the action in more detail. Shape, Texture and Motion features are extracted using HOG, LBP and DWT technique respectively. STIP is used to extract salient points from the frame. Feature selection is done by applying PCA, which helps in retaining the features carrying maximum information. Concatenation of 1st moment of Shape, Texture and Motion feature represents final feature descriptor for each frame. As compared to established methods, method proposed in this work demonstrated better performance in terms of recognition accuracy. Recognition accuracy achieved using new method, STSTM is much higher than achieved using PCA+ICA+LDA with HMM in [21]

on Weizmann dataset and it is at par with performance achieved by other approaches. In case of KTH dataset, performance of STSTM is better than WS_HAR[21], Silhouette-based pose estimation [22] and Silhouette-based SAX Shapes [23]. Comparable accuracy is achieved for UCF sports action dataset.

Table 1. Comparative results on Weizmann Dataset

State of art methods	Recognition Rate UCF	Recognition Rate Weizmann	Recognition Rate KTH
	Sports		
Carvajal et al.[20]	--	96.7%	--
Cheng et al. (Chi squared kernel) [19]	--	97.1	--
Cheng et al.(Intersection Kernel) [19]	--	96.9	--
Chaaaraoui et al.[22]	--	--	92.77%
Junejo et al.[23]	--	--	89%
Chivers et al.[24]	--	--	97%
Siddiqi et al.[21]	--	80.33%	81%
Yi, Y et al.[25]	90	--	--
Wang et al.[26]	92	--	--
Weng et al.[27]	92.8	--	--
Cho et al.[28]	89.7	--	--
This Work	93	96.4%	95.34%

References

- [1]Yuan, Yuan, Lei Qi, and Xiaoqiang Lu., 2016. Action recognition by joint learning. Image and Vision Computing Vo. 55, Part 2, pp. 77-85, November 2016.
- [2] Yi, Yun, Hanli Wang, and Bowen Zhang., 2017. Learning correlations for human action recognition in videos. Multimedia Tools and Applications, Vol 76, no. 18, pp 18891–18913, September 2017.
- [3]Shamsipour, Ghazal, JamshidShanbehzadeh, and Hossein Sarrafzadeh., 2017. Human action recognition by conceptual features. In Proc. The International MultiConference of Engineers and Computer Scientists. Vol. 1. March 2017.
- [4]Kar, Amlan, Nishant Rai, Karan Sikka, and Gaurav Sharma.,2016. AdaScan: Adaptive Scan Pooling in Deep Convolutional Neural Networks for Human Action Recognition in Videos. arXiv preprint arXiv:1611.08240 (2016).
- [5]Al-Berry, Maryam N., Hala M. Ebied, Ashraf S. Hussein, and Mohammed F. Tolba. 2014. Human action recognition via multi-scale 3D stationary wavelet analysis. In Proc 14th International Conference on Hybrid Intelligent Systems (HIS), 2014, pp. 254-259. IEEE, 2014.
- [6]Al-Berry, Maryam N., et al. 2015. Weighted Directional 3D Stationary Wavelet-based Action Classification. Egyptian Computer Science Journal 39.2 (2015).
- [7]Al-Berry, Maryam N., et al., 2016. Action Classification Using Weighted Directional Wavelet LBP Histograms. In Proc The 1st International Conference on Advanced Intelligent System and Informatics (AIS2015), November 28-30, 2015, BeniSuef, Egypt. Springer International Publishing, 2016.
- [8]Aryanfar, Alihossein, et al., 2015. Multi-View Human Action Recognition Using Wavelet Data Reduction and Multi-Class Classification. Procedia Computer Science, Vol. 62, pp 585-592, 2015.
- [9]Thanikachalam, V., and K. K. Thyagarajan., 2015. Human Action Recognition by Employing DWT and Texture. In Proc Artificial Intelligence and Evolutionary Algorithms in Engineering Systems. Springer, New Delhi, pp 309-316, 2015.
- [10]Siddiqi, Muhammad Hameed, et al., 2014. Video-based human activity recognition using multilevel wavelet decomposition and stepwise linear discriminant analysis. Sensors 14.4 (2014), pp 6370-6392.
- [11]Rahman, Saimunur, John See, and Chiung Ching Ho., 2015 Action recognition in low quality videos by jointly using shape, motion and texture features. In Signal and Image Processing Applications (ICSIPA), 2015 IEEE International Conference on, pp. 83-88. IEEE, 2015.
- [12]Wu, Qiuxia, et al.,2013. Realistic human action recognition with multimodal feature selection and fusion. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 43.4 (2013): 875-885.
- [13]Huang, Di, et al.,2014. HSOG: a novel local image descriptor based on histograms of the second-order gradients. IEEE Transactions on Image Processing 23.11 (2014): 4680-4695.

- [14]Ojala, Timo, Matti Pietikäinen, and Topi Mäenpää.,2000. Gray scale and rotation invariant texture classification with local binary patterns. In Proc European Conference on Computer Vision, pp. 404-420. Springer, Berlin, Heidelberg.
- [15]O'Hara, Stephen, and Bruce A. Draper., 2011. Introduction to the bag of features paradigm for image classification and retrieval. arXiv preprint arXiv:1101.3354 (2011).
- [16] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, 2005. Behavior recognition via sparse spatio-temporal features. In Proc. IEEE Int. Work-shop Vis. Surveill. Perform. Eval. Tracking Surveill., Jun. 2005, pp. 65-72.
- [17]Dalal, Navneet, and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005. Vol. 1. IEEE, 2005.
- [18]Nguyen, Thanh Phuong, et al., 2013. Revisiting lbp-based texture models for human action recognition. Iberoamerican Congress on Pattern Recognition. Springer, Berlin, Heidelberg, 2013.
- [19]Cheng, Shilei, et al., 2016. Action Recognition Based on Spatio-temporal Log-Euclidean Covariance Matrix. International Journal of Signal Processing, Image Processing and Pattern Recognition 9.2 (2016): 95-106.
- [20]Carvajal, Johanna, et al., 2016 Comparative Evaluation of Action Recognition Methods via Riemannian Manifolds, Fisher Vectors and GMMs: Ideal and Challenging Conditions. arXiv preprint arXiv:1602.01599 (2016).
- [21]Siddiqi, Muhammad Hameed, et al., 2014. Video-based human activity recognition using multilevel wavelet decomposition and stepwise linear discriminant analysis. Sensors 14.4 (2014): 6370-6392.
- [22]Chaaaraoui, Alexandros Andre, Pau Climent-Pérez, and Francisco Flórez-Revuelta.,2013. Silhouette-based human action recognition using sequences of key poses. Pattern Recognition Letters 34.15 (2013): 1799-1807.
- [23]Junejo, Imran N., Khurum Nazir Junejo, and Zaher Al Aghbari.,2014. Silhouette-based human action recognition using SAX-Shapes. The Visual Computer 30.3 (2014): 259-269.
- [24]Chivers, Daniel Stephen.,2012. Human Action Recognition by Principal Component Analysis of Motion Curves. PhD diss., Wright State University
- [25] Yi, Y. and Lin, Y., 2013. Human action recognition with salient trajectories. Signal Processing, 93(11), pp.2932-2941.
- [26] Wang, X. and Qi, C., 2016. Action recognition using edge trajectories and motion acceleration descriptor. Machine Vision and Applications, 27(6), pp.861-875.
- [27] Weng, Z. and Guan, Y., 2018. Action recognition using length-variable edge trajectory and spatio-temporal motion skeleton descriptor. EURASIP Journal on Image and Video Processing, 2018(1), p.8.
- [28] Cho, J., Lee, M., Chang, H.J. and Oh, S., 2014. Robust action recognition using local motion and group sparsity. Pattern Recognition, 47(5), pp.1813-1825.

Authors Biography



Ms. Aditi S. Jahagirdar, received her Bachelor's degree in 1993 in Electronics & Telecommunication engineering and Master's degree in Electronics Engineering in 1999 from University of Pune, Maharashtra, India. She is currently working as Assistant Professor in Department of Information Technology at MIT College of Engineering, Pune. She is having total 21 years of teaching experience. She is currently working towards her Ph.D. degree in Electronics Engineering in Savitribai Phule University, Pune, India. She is life member of 'Computer Society of India' and 'Institution of Electronics and Telecommunication Engineers'. Her research interests include computer vision, Image processing, Video processing and pattern recognition.



M.S. Nagmode, is Currently working as Professor and HOD in Government College of Engineering and Research, Avasari, Pune, India. He is Professor in Department of Electronics and Telecommunication Engineering Department and is having teaching experience of 20 years. He has published 41 research papers in various journals and conferences. His areas of interest include Image processing, Signal processing and VLSI design. He was awarded Ph.D. degree by University of Pune, India in 2009 in Electronics and Telecommunication Engineering in the domain of Image Processing. He received his Bachelor's degree in Electronics Engineering in 1996 and Master's degree in Electronics Engineering in 2000 from University of Pune India.

