# Efficient Sparse Representation based Action Recognition in video

Ushapreethi P, Lakshmi Priya G G

***Abstract:*** ***Human Action Recognition (HAR) is an interesting and helpful topic in various real-life applications such as surveillance based security system, computer vision and robotics. The selected features and feature representation methods, classification algorithms decides the accuracy of the HAR systems. A new feature called, Skeletonized STIP (Spatio Temporal Interest Points) is identified and used in this work. The skeletonization on the action video's foreground frames are performed and the new feature is generated as STIP values of the skeleton frame sequence. Then the feature set is used for initial dictionary construction in sparse coding. The data for action recognition is huge, since the feature set is represented using the sparse representation. To refine the sparse representation the max pooling method is used and the action recognition is performed using SVM classifier. The proposed approach outperforms on the benchmark datasets.***

***Keywords:*** ***Skeletonization Sparse representation, action recognition, sparse coding, sparse dictionaries, SVM classifier.***

## I. INTRODUCTION

Computer vision is a complicated task consists of various levels of sub tasks such as object detection, person identification, gesture recognition and action recognition [1-4]. The complications of computer vision increases due to illumination changes, camera motion, and camera rotation. The computer is good in doing things very fast, but its reasoning capacity is very low compared to humans. The computer has to interpret the test action video with the training data for gaining the knowledge on the sub tasks. This knowledge plays an important role in higher end systems such as robots and automated vehicles. The ability of the computer system in identifying the human actions performed in a video is known as human action recognition.

Human action recognition is a highly demanding research area because of its wider applications such as computer vision in automated systems, video surveillance based security systems in many public places such as airports, railway stations and even in our own residential circumstances. The human actions are classified into four types. They are gestures, actions, interactions and group activities. Gestures are the primary movements of the human body with meaningful information. Facial gestures and hand gestures recognition are attracting areas of research due to its applications such as medical surveillance.

The action is the combination of multiple gestures, thus the complication in identifying action is higher than the gesture identification. The gesture and action is performed by single person and the interaction is performed by two persons. The group activity is performed by more than two persons. The action of the human provides the way to identify the activity of a particular scenario such as 'leaving a bag in airport', 'crossing the railway track' and more. Automatic identification of human actions and activities can be performed by computer systems with proper training. There are several ongoing researches are based on the automatic action identification perspective. The major steps of sparse representation based action recognition is shown in Figure 1.

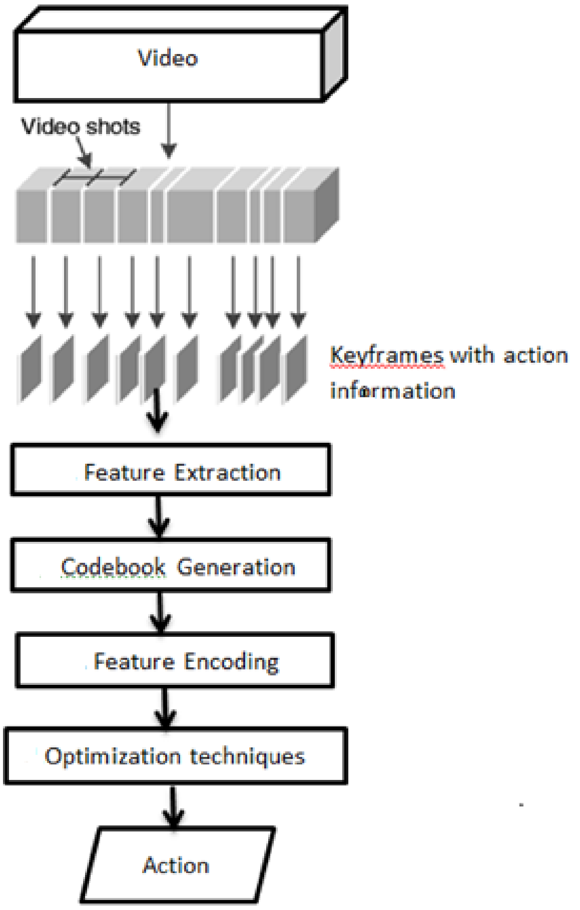

**Fig. 1 Action Recognition phases**

The sparse representation based action recognition phases are feature extraction, codebook generation, feature encoding and optimization techniques, action classification [5]. Initially, the video is divided into frames of key-frames. Then the low level features such as Histogram of Oriented Gradients (HOG) [6], Histograms of Optical Flow (HOF) [7],

Revised Manuscript Received on December 30, 2019.
* Correspondence Author

**Ushapreethi P,** School of Information Technology and Engineering, Vellore Institute of Technology, Vellore. Email: ushapreethi.p@vit.ac.in.
**Lakshmi Priya G G,** School of Information Technology and Engineering, Vellore Institute of Technology, Vellore. Email: lakshmipriya.gg@vit.ac.in.

STIP [8], Bag Of Visual Words (BoW) [9], Motion Boundary Histograms (MBH) [10], Local Binary Patterns (LBP) [11] are extracted; and known as feature extraction. The dictionary for sparse representation is constructed using the low level features; known as code book generation. The low level features are appended for improvising the dictionary. The sparse matrix is considered and the mapping between the training features and the dictionaries are done and represented in the sparse matrix; known as feature encoding. The sparse matrix is improvised with the help of some optimization techniques such as Principal Component Analysis (PCA) [12], max pooling [13] and sum pooling [14]; known as optimization. Finally the action is classified using any of the well-known classifier; known as action classification. The product of the action classification is action and used for future perusal.

## II. PROPOSED FRAMEWORK

The skeletonization process is the key process in feature extraction. The frames of the training video are extracted and for a particular action. The foreground images of the frames are extracted using Gaussian Mixture Model (GMM) [15]. The skeletonization algorithm is applied on the foreground images and the skeletons are extracted for all the frames. The STIP values for the skeleton frames are extracted and the initial dictionary is framed. The initial dictionary consist of only one action's STIP values. A training video is given and the skeletonized STIP values are extracted using the same process and the training features and the testing features are provided to the linear SVM classifier. The proposed framework is shown in the figure 2. The method performs well same as the original STIP value based method with low dimension. Then the dataset is loaded and the training and test data ratio is defined and the accuracy of the proposed method is plotted.

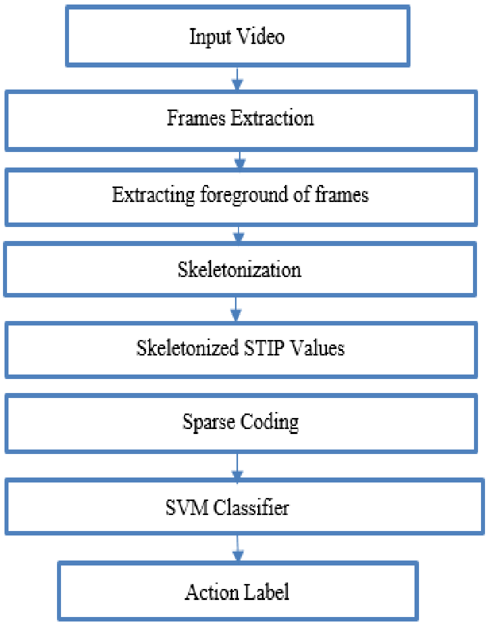

**Fig. 2 Proposed Framework**

## III. FEATURE SELECTION

The skeleton frames are considered in this work for its low dimensional value. The skeleton frames are extracted from the foreground frames using the skeletonization algorithm given below. The STIP values of extracted skeletons are identified, which provides the spatial and temporal difference measure with respect to sequence of frames. The STIP values of the skeletons are very less compared to original frames. Figure 4 shows the STIP representation of the action and the idea skeleton based representation clearly. Although the number of STIP values are reduced in Skeletonized STIP, the efficiency in finding action is good same as the original STIP. The identified low level features are used for the next phase of the action recognition, codebook generation.

**Skeletonization Algorithm**

Input: foreground F= [$P_1$, $P_2$, …, $P_n$]
Output: Skeleton S=($P_i$), iϵ1,..,n
Process:
Step 1: Remove Pixel $P_i$, if, equations1-4 are satisfied by considering $P_{i=}$ $P_1$ on the image grid shown in Figure 3.
$2 \le N(P_1) \le 6$ ------------- (1)
$M(P_1) = 1$ --------------------(2)
$P_2 . P_4 . P_6 = 0$ --------------- (3)
$P_4 . P_6 . P_8 = 0$ --------------- (4)
$N(P_1)$ is the number of neighbors of Pi in figure 3
$M(P_1)$ is the 0-1 transition in figure 3
Step 2: else if, remove Pixel $P_i$, if equations 1,2,5,6 are satisfied
$P_2 . P_4 . P_8 = 0$ **---------------- (5)**
$P_2 . P_6 . P_8 = 0$ ---------------- (6)

| | | |
|---|---|---|
| P₉ | P₂ | P₃ |
| P₈ | P₁ | P₄ |
| P₇ | P₆ | P₅ |

**Fig. 3 Mask for Skeletonization**

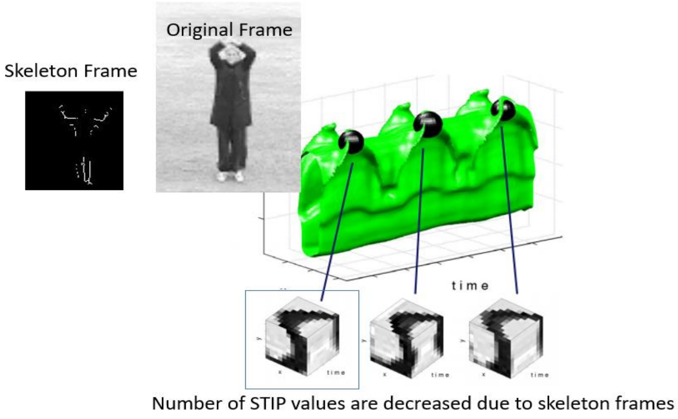

**Fig. 4 Skeletonized STIP values.**

## IV. SPARSE CODING

Sparse coding is introduced by [16] for image classification. The low level features are represented in terms of sparse matrix with the help of dictionaries. The dictionaries consists of basis column vectors called atoms. Input skeletonization STIP values are represented as x(1), x(2), …, x(m) The dictionary with bases f1, f2, …, fk is created, so that each skeletonization STIP value x can be represented as:

$$x \approx \sum_{j=1}^{k} aj\ \varnothing j \text{-----------------------} (7)$$

In (1), aj is the coefficients vector with respect to the atom feature x and dictionary atom $\varnothing j$, which is computed as $< x * \varnothing j >$. The transform coefficients vector ‘a’ varies for different features. The variation in the transform coefficient matrix can be represented using Fourier transform matrix, discrete cosines (DCT matrix) and discrete sines (DST matrix), Haar transform matrix, wavelet and wavelet packets matrices, Gabor filters. Most of the input features are represented in a few coefficients ‘M’ in a vector ‘a’. The other N − M coefficients have less contribution in representing a signal vector x ∈ $R_{M\times 1}$.

The features which are having less than certain value as their coefficient are set to zero in coding scheme. Maximum of the $a_j$’s value are zero in the sparse matrix providing “sparse representation” of the feature x. Then the sparse matrix is constructed for the whole training video frames and utilized for action recognition.

## V. ACHIEVING ACTION RECOGNITION BASED ON SPARSE REPRESENTATION

The high level features are the input for the SVM classifier. The sparse coding is the construction of sparse matrix with help of dictionary and the arriving training feature. The sparse coding problem constructed as the constrained optimization problem as below.

$$\min_{a,\varnothing} \sum_{i=1}^{m} \left( \|x^{(i)} - \sum_{j=1}^{k} a_j^{(i)} \varnothing_j\|^2 + \lambda \sum_{j=1}^{k} \|a_j^{(i)}\| \right) \text{---------} (8)$$

$\lambda \sum_{j=1}^{k} \|a_j^{(i)}\|$ is known as $l_1$ sparsity equation and can be easily solved.

Norm is used to get a scalar value from a vector or a matrix. It varies according to the optimization problem. In some cases, it is the total size or length of all vectors in a vector space or matrices. If the norm is high, the matrix value is also high. In general, the norm is represented as ||x|| where x is a vector. Formally the $l_p$ norm of x is defined as: $\|x\|_p = \sqrt[p]{(i|x|_p)}$ where $p \in \mathbb{R}$. According to the p value, the norm value varies from 0 to ∞.

The sparse representation based action recognition works have taken various methods to solve the sparse coding problem efficiently. The methods such as Alternating Direction Method of Multipliers (ADMM) [17], Lagrange Dual [18] are most famous among them. The problem has been converged using generalized Lagrangian method.
Thus the final sparse features are given to the SVM classifier and K-Means classifier for analyzing the performance of the proposed method. Support Vector Machine (SVM) classifier classifies the action better compared to K-Means classifier.

## VI. RESULTS AND DISCUSSION

The work is carried out on KTH [19] and Weizmann datasets [20]. KTH dataset consists of 600 videos on 6 different actions namely, run, walk, box, jog, one hand-wave and two hand-wave. The Weizmann dataset consists of 84 videos with 7 different actions namely run, walk, skip, jump, bend, hand-clap and hand-wave. The original frames, foreground frames and skeleton frames for KTH walk action are shown in figure 5.

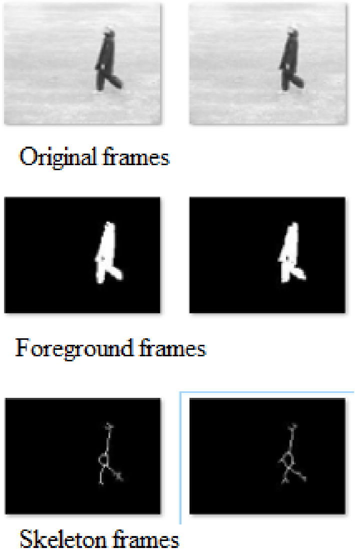

**Fig. 5 Frames, foreground and skeletons.**

The experiments are carried out with original STIP features foreground STIP features and skeleton STIP features were carried out and results are considered for analysis. The dataset is considered as 1:4 ratio on training and testing data. Table 1 and Table 3 Shows the comparison between three types of low level STIP feature and their performances. The skeleton STIP descriptor provides better results than other descriptors on both SVM and K-Means classifiers. The accuracy of the proposed method is 98.14 for KTH dataset and 96.14 for Weizmann dataset, which is obtained using SVM classifier.

**Table 1. Comparison of various methods on KTH and Weizmann dataset using SVM classifier.**

| Descriptors | Accuracy | |
|---|---|---|
| | KTH (%) | Weizmann (%) |
| **Original STIP** | 89.32 | 82.16 |
| **Foreground STIP** | 97.09 | 89.34 |
| **Skelton descriptor** | **98.17** | **96.14** |

Table 2. Comparison of Size of the descriptor.

| Number of frames | Number of STIP values for original color frame | Number of STIP values for original color frame | Number of STIP values for Skeletonized frame | Size of the feature descriptor with original frames | Size of the feature descriptor with foreground frames | Size of the feature descriptor with skeleton frames |
|---|---|---|---|---|---|---|
| 20 | 12 | 10 | 8 | 240 | 200 | 160 |
| 25 | 12 | 10 | 8 | 300 | 250 | 200 |
| 40 | 12 | 10 | 8 | 480 | 400 | 320 |
| 50 | 12 | 10 | 8 | 600 | 500 | 400 |

Table 3. Comparison of various methods on KTH and Weizmann dataset using K-Means classifier.

| Descriptors | Accuracy | |
|---|---|---|
| | KTH (%) | Weizmann (%) |
| Original STIP | 89.76 | 83.16 |
| Foreground STIP | 96.99 | 90.28 |
| Skelton descriptor | 96.24 | 95.23 |

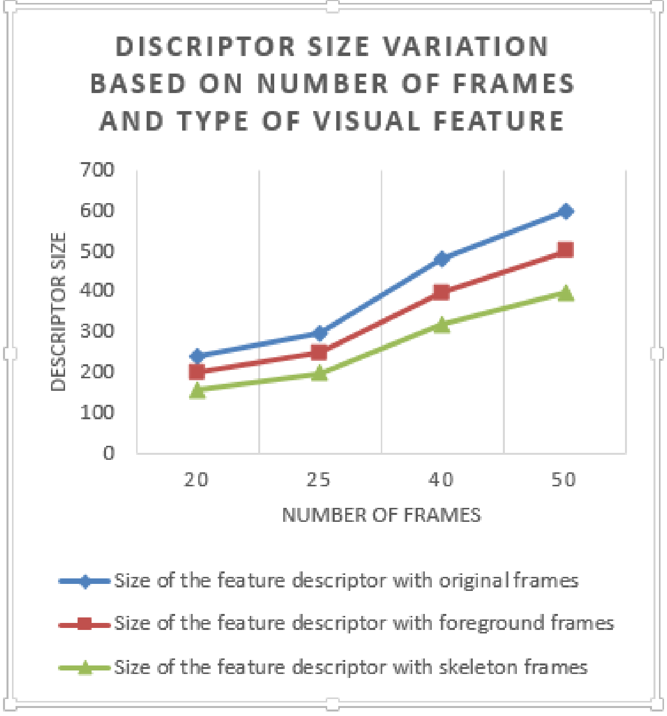

Fig 6. Comparison on descriptor size based on Table 3.

The major contribution of the work is the reduction in low level feature descriptor size. Table 2 and Figure 6 shows the variation in the size of low level feature descriptor which varies based on the frame selected (original frame, foreground frame and skeleton frame) and the number of frames taken for representing an action in the video. The original frames contains the background information, which moves along with the foreground, hence the size of the descriptor (STIP values) increases. The foreground frames contains less STIP values compared to original frame, because the foreground used in this method is GMM with smoothening. The skeleton descriptor occupies less size compared to both the other descriptors, because very less active information present in the skeleton STIP values. A major reason needs to be given on the accuracy level of the Skeletonized STIP values.

A sample of skeleton sequence is given for the reasoning behind the idea. Open the image shown in Figure 7 (comprises of 6 images) in the photo viewer, view all the images continuously by pressing the right arrow in the keyboard. The action will be recognized by you. The same logic is applied in this work with STIP values. The STIP values are sparsely represented and the sparse data is used for the action classification. However, the system is capable enough to identify the action with proper training data.

## VII. CONCLUSION AND FUTURE WORK

The proposed method achieved good accuracy with identified new skeletonized STIP features and discriminative sparse representation. The maximum sparsity is achieved using the skeletonized STIP features, which helps for efficient classification. The sparse features are better utilized by both the K-Means and SVM classifiers. The basic sparse representation is utilized in our work and the sparse coding is effectively done using Lagrange multipliers. The results of the proposed method are very good on the given datasets. The proposed framework can be applied on dynamic background video datasets in future.

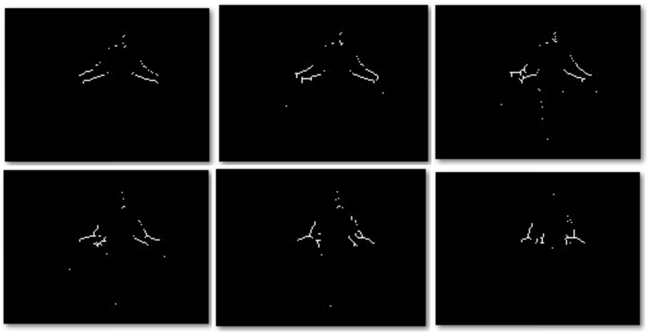

Fig. 7 skeleton frames

## REFERENCES

1. JK Aggarwal, MS Ryoo, “Human activity analysis: a review,” *ACM Computing Survey* vol.43, pp. 1–43, April 2011.
2. Liu, Y., Nie, L., Liu, L. and Rosenblum, D.S., 2016. From action to activity: sensor-based activity recognition. Neurocomputing, 181, pp.108-115.
3. Dawar, N. and Kehtarnavaz, N., 2018. Action detection and recognition in continuous action streams by deep learning-based sensing fusion. IEEE Sensors Journal, 18(23), pp.9660-9668.
4. MA Bagheri, Q Gao, S Escalera, TB Moeslund, H Ren, E Etemad, “Locality regularized group sparse coding for action recognition,” Computer Vision and Image Understanding, vol. 158, pp. 106–114, May 2017.

5. Z Gao, SH Li, YJ Zhu, C Wang, H Zhang, “Collaborative Sparse Representation Leaning Model for RGBD Action Recognition,” Journal of Visual Communication and Image Representation. March 2017.
6. Ji, Xiaopeng, et al. "The spatial Laplacian and temporal energy pyramid representation for human action recognition using depth sequences." Knowledge-Based Systems 122 (2017): 64-74.
7. Gao, Chenqiang, et al. "Infar dataset: Infrared action recognition at different times." Neurocomputing 212 (2016): 36-47.
8. Li, Lingqiao, et al. "Learning a Discriminative Feature Descriptor with Sparse Coding for Action Recognition." 2018 17th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES). IEEE, 2018.
9. Abdulmunem, Ashwan, Yu-Kun Lai, and Xianfang Sun. "Saliency guided local and global descriptors for effective action recognition." Computational Visual Media 2.1 (2016): 97-106.
10. Varol, Gül, and Albert Ali Salah. "Efficient large-scale action recognition in videos using extreme learning machines." Expert Systems with Applications 42.21 (2015): 8274-8282.
11. Baumann, Florian, et al. "Recognizing human actions using novel space-time volume binary patterns." Neurocomputing 173 (2016): 54-63.
12. Wen, Jiajun, et al. "The L2, 1-norm-based unsupervised optimal feature selection with applications to action recognition." Pattern Recognition 60 (2016): 515-530.
13. Liu, Huaping, Mingyi Yuan, and Fuchun Sun. "RGB-D action recognition using linear coding." Neurocomputing 149 (2015): 79-85.
14. Ji, Xiaopeng, et al. "The spatial Laplacian and temporal energy pyramid representation for human action recognition using depth sequences." Knowledge-Based Systems 122 (2017): 64-74.
15. Chen, Mingliang, et al. "Spatiotemporal GMM for background subtraction with superpixel hierarchy." IEEE transactions on pattern analysis and machine intelligence 40.6 (2017): 1518-1525.
16. Yang, Jianchao, et al. "Linear spatial pyramid matching using sparse coding for image classification." 2009 IEEE Conference on computer vision and pattern recognition. IEEE, 2009.
17. Han, D. and Yuan, X., 2012. A note on the alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, *155*(1), pp.227-238.
18. Liu, W., Zha, Z.J., Wang, Y., Lu, K. and Tao, D., 2016. $ p $-Laplacian regularized sparse coding for human activity recognition. *IEEE Transactions on Industrial Electronics*, *63*(8), pp.5120-5129.
19. Schuldt, C., Laptev, I., Caputo, B., 2004. Recognizing human actions: A local svm approach. In Proc. ICPR, pp. 32–36.
20. Gorelick, L., Blank, M., Shechtman, E., Irani, M. and Basri, R., 2007. Actions as space-time shapes. IEEE transactions on pattern analysis and machine intelligence, 29(12), pp.2247-2253.

## AUTHORS PROFILE

**P.Ushapreethi** graduated in BE (Computer Science and Engineering) degree from Anna University in the year 2009 and ME (Multimedia Technology) from Anna University in the year 2011. She had 3 years of industrial experience and she has been teaching for the past five years and presently working as Assistant Professor in School of Information Technology and Engineering at Vellore Institute of Technology, Vellore, India. She is currently doing research and her research areas include video segmentation methods, video content analysis, and feature encoding techniques.

**Lakshmi Priya G. G.** is currently working as Associate Professor in School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India. She received the M.C.A. and M.E. degrees in 2004 and 2007, respectively and the Ph.D. degree from the National Institute of Technology at Tiruchirappalli, India in 2014. Her research interests include temporal video segmentation, content-based video retrieval, and big data - video analysis.