International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Shift Invariant Dictionary Learning for Human Action Recognition

Ushapreethi P, Lakshmi Priya G G

Abstract: ***Sparse representation is an emerging topic among researchers. The method to represent the huge volume of dense data as sparse data is much needed for various fields such as classification, compression and signal denoising. The base of the sparse representation is dictionary learning. In most of the dictionary learning approaches, the dictionary is learnt based on the input training signals which consumes more time. To solve this issue, the shift-invariant dictionary is used for action recognition in this work. Shift-Invariant Dictionary (SID) is that the dictionary is constructed in the initial stage with shift-invariance of initial atoms. The advantage of the proposed SID based action recognition method is that it requires minimum training time and achieves highest accuracy.***

Keywords: Sparse representation, action recognition, sparse coding, shift invariant dictionaries, SVM classifier

I. INTRODUCTION

Sparse representation is the efficient method for representing the huge data as an analyzable data, then the data can be used for enormous real-time applications such as medical systems, security systems, transportation systems and more. The action recognition is one of the famous topic used in real-life applications like surveillance systems, robotics and computer vision [1-4]. The link between these two giant fields are elaborated in this paper as follows. The data used in the action recognition system is the extracted features of all the frames of the video. The size of the feature is huge and it needs to be represented as sparse feature. Actions of the human in a video is recognized by applying the following procedural steps. The first step is to identify the visual (or) low level features such as color similarities, intensity similarities of the video frames. Color and intensity similarity among the sequence of frames are measured as the standard features such as Bag Of Visual Words (BoVW), Histogram of Oriented Gradients (HOG), Histograms of Optical Flow (HOF), Spatio Temporal Interest Points (STIP), Motion Boundary Histograms (MBH), Local Binary Patterns (LBP) and its variations. The resultant data of step 1 is dense feature the second step is to find the sparse feature from the dense feature.

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.

Let's consider the low level feature is X, and the sparse representation of X is given by

$$\|X - DW\| \qquad (1)$$

where D is the dictionary of size NXK and W is the sparse matrix. The dictionary construction is based on low-level features and shift-invariance [5-7].

The shift-invariant dictionary is represented as D^s, the sparse representation of X becomes,

$$\|X - D^s W\| \qquad (2)$$

The column of the dictionary is tend to be shifted for given number of times to get an over complete dictionary. The shift-invariant dictionary D^s can be diagrammatically represented as a matrix shown below.

$$\begin{bmatrix} *0 & *n & \vdots & \dots & +0 & +n & \vdots \\ *1 & *0 & *n & \dots & +1 & +0 & +n \\ *2 & *1 & *0 & \dots & +2 & +1 & +0 \\ \vdots & *2 & *1 & \dots & \vdots & +2 & +1 \\ \vdots & \vdots & *2 & \dots & \vdots & \vdots & +2 \\ *n & \vdots & \vdots & \dots & +n & \vdots & \vdots \end{bmatrix}$$

Consider '*i' is the first initial atom consider the dense feature of action 'Run' and '+i' is another initial atom of the dictionary, here the action is walk. They are circularly shifted in column-wise and the initial dictionary is constructed.
There are several types of dictionaries constructed and used for sparse representation. Some of them are fixed dictionaries [8-10], double sparse dictionaries [11-13], and shift-invariant dictionaries. The fixed dictionary $D = [D_1, D_2, \dots, D_n]$, contains n columns (atoms). Each atom is the basic feature of anyone frame in the action video and the dictionary is constructed using various unique other atoms from the training data. The double dictionary is the sparse dictionary and it is represented as $D = \emptyset A$, where $\emptyset$ is the sparse vector and A is the minimal dictionary. Double sparsity reduces the space complexity of a classification system however, it increases the time complexity and there is no specific method to solve the double dictionary problem without accuracy compromisation. The advantages of shift invariant dictionaries are twofold. First, shift invariant dictionaries consume very less time for the dictionary construction and dictionary learning. Second the shift invariant dictionary method is the efficient method for dynamically varying video datasets. Because the SID takes only one frame and performing shift invariance for other atoms. So, the training time and testing time are reduced efficiently.

II. PROPOSED FRAMEWORK

The major functions of the action recognition are aligned under three divisions in the proposed framework. They are low level feature extraction, high level feature extraction and action classification. In the first division, the input video frames (action frames) are extracted for a specific action video and stored. The STIP values for each action frames are extracted and given to the second division. The second division is the high level feature extraction. The large amount of low level features needs to be represented in a compact representation to avoid the complications in the third division. Thus the SID is constructed and the sparse coding is performed for getting the sparse features. The third division gets the sparsely represented high level feature and they are used by the classifier. Both the first and second divisions are performed on a set of training videos and the action labels are assigned according to the actions. The classifier performs the action classification on testing data based on the training data and assigned action labels assigned during the training phase. Figure 1 shows the proposed framework with all the three divisions.

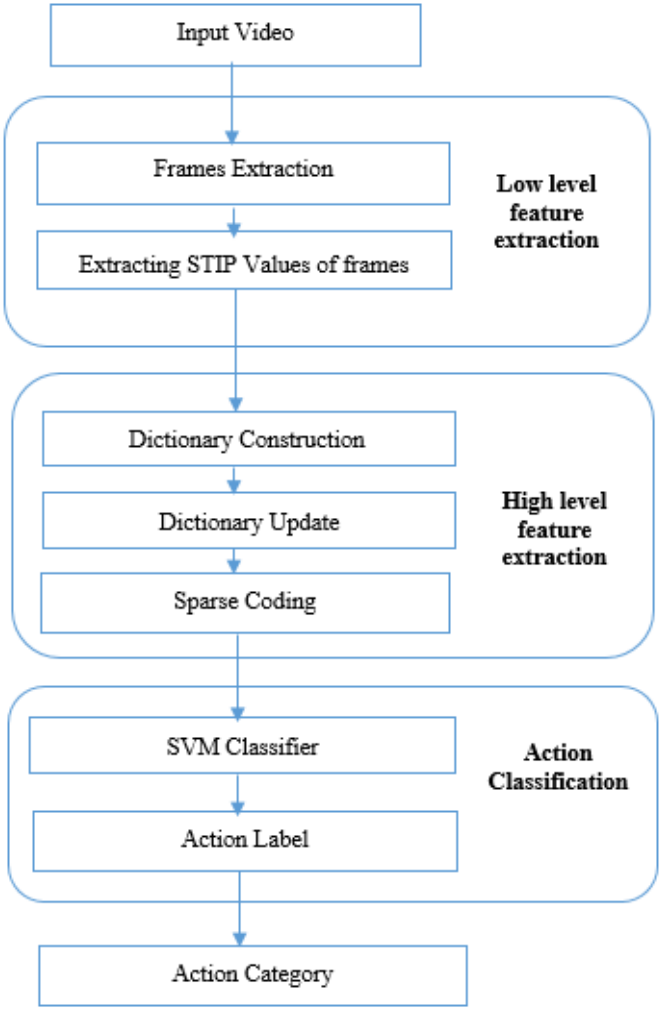

Fig. 2 Proposed Framework

III. FEATURE SELECTION

STIP values of the video frames are considered as low level feature in the proposed work. STIP values are the difference between two sequence frame's basic corner points according to the time. The varying spatial and temporal interest point values of the video frames are expressed as

$$(\nabla F)^T V + F_t = 0 \qquad (3)$$

Where, $V = (V_X, V_Y)^T$ is known as optical flow, F_t is representing the frame with respect to time t and $\nabla I = (F_X, F_Y)^T$ is the gradient. The derivatives of the above equation such as moment matrices, Gaussian kernel are used to represent the feature vector more concise and efficient. The STIP values are the 3-dimensional values with respect to x, y and t. x and y are the spatial pixel coordinates and t is the time. The basic representation pictorial of STIP values are shown in Figure 2.

STIP Algorithm [14]

Input: Frames= [F_1, F_2, …, F_n]
Output: STIP values of each frames
Process:
Step 1: Find the sum of squared differences (SSD) between two frames of the video $f(F_{x,y,t})$ based on time t
$(\nabla F)^T V + F_t = 0$
Step 2: Find the second order moment matrix

$$\begin{pmatrix} F_x^2 & F_xF_y \\ F_xF_y & F_y^2 \end{pmatrix} V = -\begin{pmatrix} F_xF_t \\ F_yF_t \end{pmatrix}$$

Step 3: Apply spatio-temporal Gaussian kernel matrix for smoothing

$$\mu = g(.,\sigma_i^2,\tau_i^2) = \begin{pmatrix} F_x^2 & F_xF_y & F_xF_t \\ F_xF_y & F_y^2 & F_xF_t \\ F_xF_t & F_xF_t & F_t^2 \end{pmatrix}$$

Step 4: Apply l_2 norm on identified results to get normalized STIP values

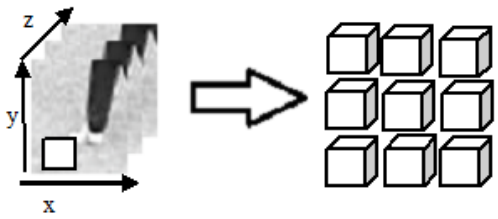

Fig. 2 Spatio Temporal Interest Points

IV. SPARSE CODING

Sparse coding is the concept introduced for image classification [15]. The low level features are used as columns (atoms) in the dictionary D in dictionary construction phase. Consider the dimensions of the dictionary D is NXK, the input signal dimension is NXP and the sparse matrix dimension is KXP. The incoming training features are known as input signals X. The input signals are compared with the dictionary and the maximum similar features are represented in the sparse matrix A with its coefficient value. Figure 3 shows the diagrammatic representation is basic sparse coding. The atoms 3 and 10 are the matching atoms in the D for the input signal 3. Thus all the input signals in X are represented as sparse values in A.

The input signal X can be represented as:

$$X = \sum_{i=1}^{k} x_i \alpha_i \qquad (4)$$

where, α_i is the dictionary atom and x_i is the coefficient vector with respect to the dictionary atom, which can be computed as $<x_i * \alpha_i>$. The transform coefficients vector 'α' varies for different features. The representation of the input signal can be based on the off the shelf transforms of input vectors such as Fourier transform matrix, discrete cosines (DCT matrix) and discrete sines (DST matrix), Haar transform matrix, wavelet matrices, Gabor filters. But the training features are considered in the image processing works. The input signal with size NX1 is represented with few values in the sparse matrix. The minimum value coefficients are considered as 'zero' value coefficients in the coding scheme.

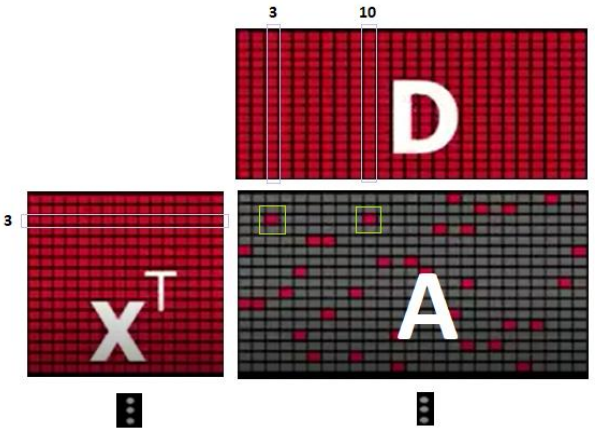

Fig. 4. Sparse Coding

V. ACHIEVING ACTION RECOGNITION BASED ON SPARSE REPRESENTATION

Our goal is to minimize the difference between the original feature and the sparse feature to get the correct action label from the classifier. The action recognition based on sparse representation is formulated as a constrained optimization problem as follows.

$$W, D = \underset{W,D}{\operatorname{argmin}} \|X - DW\|_F^2, s.t\, W \text{ is sparse} \qquad (5)$$

where, $\|.\|_F^2$ is the Frobenius norm. By solving the constrained optimization problem, the sparse features for the STIP values can be obtained.

The problem can be solved using two steps. The first step is to fix the dictionary D and to find the sparse matrix W. Consider $X = [x_0, x_1, \ldots, x_{L-1}]$ is the input signal vector and $W = [w_0, w_1, \ldots, w_{L-1}]$ is the sparse vector in the sparse matrix. The problem for the first step can be formulated as

$$w_i = \underset{w}{\operatorname{argmin}} \|x_i - Dw\|_F^2, s.t\, \|w\|_0 \le S, 0 \le i \le L-1 \qquad (6)$$

The second step is to fix the sparse matrix W and to update the dictionary D. The second step can be formulated as

$$D = \underset{D}{\operatorname{argmin}} \|X - DW\|_F^2, s.t\, vec(D) = vec(D^S) \qquad (7)$$

The first step is known as sparse coding and second step is known as dictionary update. The sparse coding problem is solved efficiently using generalized lagrangian multiplier method. The first order lagrangian derivatives are identified with the sparse (learning rate) parameter for equation 6 until the convergence point is reached. Thus the high level sparse features are identified and given to the SVM classifier. The process from extracting the low level feature to high level features is represented as diagram given in Figure 4.

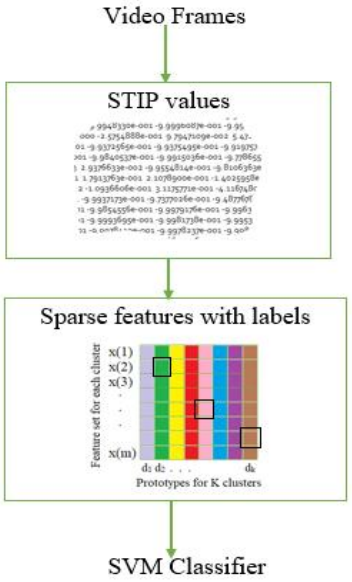

Fig. 4. video to sparse features conversion.

VI. RESULTS AND DISCUSSION

The KTH [16] and Weizmann [17] datasets are considered for experiments. The KTH dataset consists of 6 actions (run, walk, box, jog, one hand-wave and two hand-wave) and 100 videos for each action has been taken, and the size of the dataset is 600 videos. The Weizmann dataset consists of 84 videos in total for representing 7 actions (run, walk, skip, jump, bend, hand-clap and hand-wave). The initial 50 frames of each video are considered for experiments based on two major variations. The first variation is the type of feature considered. The experiments are carried out with original STIP values foreground STIP values and results are considered for analysis. The second variation is the type of dictionary. SID and the normal dictionaries are considered for evaluation. The dataset is split into training data and testing data with the ration 1:4.

Table 1. Accuracy comparison of various methods on KTH and Weizmann dataset.

Descriptors	Accuracy	
	KTH (%)	Weizmann (%)
Original STIP	89.32	82.16
Foreground STIP	97.09	89.34
Original STIP + Sparse coding (normal dictionary)	**98.17**	96.14
Foreground STIP + Sparse coding (normal dictionary)	97.82	97.09
Original STIP + Sparse coding (SID)	98.01	**97.11**
Foreground STIP + Sparse coding (SID)	98.12	96.23

Table 1 shows the accuracy comparison of different feature descriptors (feature representation methods) on KTH and Weizmann datasets. Table 2 shows the execution time comparison of various feature descriptors on both KTH and Weizmann datasets. The maximum accuracy for KTH dataset is achieved when the original frame STIP values are sparsely represented using normal dictionary. The maximum accuracy for Weizmann dataset is achieved for the proposed method, combining the original STIP with SID based sparse representation and the maximum accuracy values are highlighted in the table 1.

Table 2. Execution time comparison of various methods on KTH and Weizmann dataset.

Descriptors	Execution Time	
	KTH (sec)	Weizmann (sec)
Original STIP	1602	520
Foreground STIP	1500	480
Original STIP + Sparse coding (normal dictionary)	1720	610
Foreground STIP + Sparse coding (normal dictionary)	1580	510
Original STIP + Sparse coding (SID)	1100	320
Foreground STIP + Sparse coding (SID)	**1024**	**310**

The major achievement of the proposed work is the reduction in the execution time. Table 2 and Figure 5 shows the variation in the execution time which varies based on the type of frame selected (original frame, foreground frame) and the descriptor for representing an action in the video. The execution time is very less for the proposed SID based sparse representation taken on STIP values of the video frames. Both the KTH dataset and Weizmann dataset are executed by the system with less execution time. The accuracy of the proposed work is also achieved for Weizmann dataset. The training time for the proposed system is very less because of the SID, thus the system is well suitable for the environment with dynamic video inputs such as abnormal activity identification of patients, abnormal activities of humans in public places.

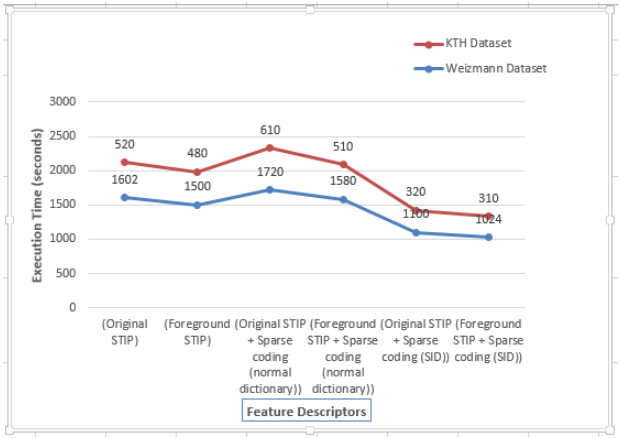

Fig 5. Comparison on execution time based on Table 2.

VII. CONCLUSION AND FUTURE WORK

In this paper, a new framework for human action recognition is proposed based on the STIP values of video frames and SID based sparse representation. STIP values provide the best spatial and temporal differences to find the movements in the video (actions). Sparse representation is used to reduce the size of low level features (STIP) and to provide high level sparse feature to the classifier. The SID concept is used to reduce the dictionary construction time in sparse representation. Our experiment results shows that the goal of the proposed work is achieved with very low range accuracy compromisation. The system used the benchmark datasets and the accuracy and execution time are notified. However, the system is well suitable for real time dynamic videos.

REFERENCES

1. Kjellström, H., Romero, J. and Kragić, D., 2011. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, *115*(1), pp.81-90.
2. Zhu, G., Xu, C., Huang, Q., Gao, W. and Xing, L., 2006, October. Player action recognition in broadcast tennis video with applications to semantic analysis of sports game. In *Proceedings of the 14th ACM international conference on Multimedia* (pp. 431-440). ACM.
3. JK Aggarwal, MS Ryoo, "Human activity analysis: a review," *ACM Computing Survey* vol.43, pp. 1–43, April 2011.
4. Chaaraoui, A.A., Climent-Pérez, P. and Flórez-Revuelta, F., 2013. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters*, *34*(15), pp.1799-1807.
5. Zhou, H., Chen, J., Dong, G. and Wang, R., 2016. Detection and diagnosis of bearing faults using shift-invariant dictionary learning and hidden Markov model. *Mechanical systems and signal processing*, *72*, pp.65-79.
6. Yang, B., Liu, R. and Chen, X., 2017. Fault diagnosis for a wind turbine generator bearing via sparse representation and shift-invariant K-SVD. *IEEE Transactions on Industrial Informatics*, *13*(3), pp.1321-1331.
7. Eldar, Y.C., 2009. Uncertainty relations for shift-invariant analog signals. *IEEE transactions on Information Theory*, *55*(12), pp.5742-5757.
8. Wang, H., Yuan, C., Hu, W. and Sun, C., 2012. Supervised class-specific dictionary learning for sparse modeling in action recognition. *Pattern Recognition*, *45*(11), pp.3902-3911.
9. Gao, Z., Zhang, H., Xu, G.P., Xue, Y.B. and Hauptmann, A.G., 2015. Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition. *Signal Processing*, *112*, pp.83-97.
10. Zheng, J., Jiang, Z. and Chellappa, R., 2016. Cross-view action recognition via transferable dictionary learning. *IEEE Transactions on Image Processing*, *25*(6), pp.2542-2556.
11. Chen, P., Jiao, L., Liu, F., Zhao, J., Zhao, Z. and Liu, S., 2017. Semi-supervised double sparse graphs based discriminant analysis for dimensionality reduction. *Pattern Recognition*, *61*, pp.361-378.
12. Abavisani, M., Joneidi, M., Rezaeifar, S. and Shokouhi, S.B., 2015, December. A robust sparse representation based face recognition system for smartphones. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (pp. 1-6). IEEE.
13. Zong, Y., Zheng, W., Cui, Z. and Li, Q., 2016. Double sparse learning model for speech emotion recognition. *Electronics Letters*, *52*(16), pp.1410-1412.
14. Yan, X. and Luo, Y., 2012. Recognizing human actions using a new descriptor based on spatial–temporal interest points and weighted-output classifier. *Neurocomputing*, *87*, pp.51-61.
15. 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.
16. Schuldt, C., Laptev, I., Caputo, B., 2004. Recognizing human actions: A local svm approach. In Proc. ICPR, pp. 32–36.
17. 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.