I.J. Image, Graphics and Signal Processing, **2019, 9, 34-43**
Published Online September 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2019.09.04

Human Action Recognition Using Modified Bag of Visual Word based on Spectral Perception

Om Mishra
Department of Electronics & Communication, Delhi Technological University, Delhi-110042, India
Email: om.mishra83@gmail.com

Rajiv Kapoor
Department of Electronics & Communication, Delhi Technological University, Delhi-110042, India
Email: rajivkapoor@dce.ac.in

M.M. Tripathi
Department of Electrical Engineering, Delhi Technological University, Delhi-110042, India
Email: mmtripathi@dce.ac.in

Received: 02 May 2019; Accepted: 25 June 2019; Published: 08 September 2019

Abstract—Human action recognition has a very vast application such as security, patient care, etc. Background cluttering, appearance change due to variation in viewpoint and occlusion are the prominent hurdles that can reduce the recognition rate significantly. Methodologies based on Bag-of-visual-words are very popular because they do not require accurate background subtraction. But the main disadvantage with these methods is that they do not retain the geometrical structural information of the clusters that they form. As a result, they show intra-class mismatching. Furthermore, these methods are very sensitive to noise. Addition of noise in the cluster also results in the misclassification of the action. To overcome these problems we proposed a new approach based on modified Bag-of-visual-word. Proposed methodology retains the geometrical structural information of the cluster based on the calculation of contextual distance among the points of the cluster. Normally contextual distance based on Euclidean measure cannot deal with the noise but in the proposed methodology contextual distance is calculated on the basis of a difference between the contributions of cluster points to maintain its geometrical structure. Later directed graphs of all clusters are formed and these directed graphs are described by the Laplacian. Then the feature vectors representing Laplacian are fed to the Radial Basis Function based Support Vector Machine (RBF-SVM) classifier.

Index Terms—Bag-of-visual-word, Contextual distance, Directed graph, Laplacian of directed graph, SVM.

I. INTRODUCTION

Human action recognition from a video is the most promising research area of computer vision and pattern analysis. It has vast applications including surveillance, human-computer interaction, video content analysis, patient care, etc. [1, 2]. Background cluttering, noise, variation in viewpoint, variation in illumination, occlusion are the major challenges for human action recognition in a video. The actions are described by the global and local feature descriptors. The global feature descriptor gives shape as well as the motion information of the person who is performing an action. These are not invariant to the appearance change, shape changes, etc. Whereas local descriptor such as spatio-temporal points shows invariance in appearance change, background change, etc. Local features based on Bag-of-visual–words have been very popular among the researchers. The disadvantage of these methods is that they do not retain the structural information of the clusters formed in Bag-of Visual words based methodology. The proposed method overcomes the limitation of the Bag-of-visual-word methodology. The proposed methodology contributed in the following manner:

1. The contextual distance among the points of a cluster is calculated on the basis of the difference between the contributions of points to maintain the geometrical structure of the cluster, unlike the contextual distance calculated on basis of Euclidean measure among the points in a cluster.
2. The contextual distance among any two points of the contextual set of a cluster, calculated as mentioned above may show asymmetry. So to deal with this asymmetry issue, we used the concept of the directed graph. Thus, a cluster is represented using contextual distance and directed graph together in a robust way.

The rest of the paper is discussed as follows: literature review has been discussed in section 2, section 3 deals with the proposed methodology. Section 4 deals with the experimental results. In this section, the proposed

methodology is applied to challenging public datasets such as KTH, Ballet, and IXMAS. The accuracy of the proposed methodology and its comparison with other state-of-the-art methodologies are also discussed in this section. The conclusion part is discussed in section 5.

II. Literature Review

Many methodologies have been proposed by researchers to describe an action. Global features [3, 4, 5, 6, 7 and 8] require the detection of the human body whose action is to be recognized. Background subtraction or human body tracking is used to represent the human body. These methodologies used two-dimensional template matching. The silhouette of the human body detected from the frames is represented globally. The descriptors like Hu moments, Radon transform, etc. are used to describe these actions. Actions are also represented as spatiotemporal volumes [9, 10, 11 and 12]. The silhouettes are stacked together on a temporal axis to make spatiotemporal volume. The disadvantage of these methodologies is that it requires accurate background subtraction to create a 2-D template and spatiotemporal volume. These global features can only give shape information but fail to provide motion information. Some global features based on the optical flow [13, 14] are used to eliminate this problem but the disadvantage of this method is that they are very sensitive to noise.

Recently, local feature descriptions [15, 16, 17, 18, 19, 20, 21, 22, 23, 24 and 26] are used more often. These methodologies do not require accurate background subtraction. They also show invariance against the viewpoint and appearance of the object. This is represented in local patches formed around spatiotemporal interest points. Local spatiotemporal descriptors are based on the bag-of-words model. HOG/HOF, HOG3D, SURF, and MoSIFT are some major descriptors [27-29]. The main disadvantage with these methodologies is that it cannot hold the shape information of the clusters. This results in the same description for different actions.

In recent years, deep learning techniques are also getting famous in researchers for human action recognition [30, 31]. These techniques are completely automated. They require large datasets and the efficiency of these methods is based on the design of networks. These networks are made up of a huge number of data and parameters that makes the structure very complex. So it takes more time to train a model. Moreover, less availability of large datasets makes this method less generalized for action recognition. Researchers are working on deep learning techniques to overcome these problems**.**

Among all these above-discussed methodologies, spatiotemporal features points-based bag-of-word are very popular because they do not require important preprocessing like background subtraction, tracking of an object, etc. This makes it robust to the challenges like background cluttering, variation in viewpoint and illumination change. The major disadvantage of this methodology is that it cannot depict the structural relationship among spatiotemporal feature points. Due to this problem, the spatiotemporal features based on a bag-of-word methodology are not able to differentiate the similar type of actions often. Fig.1. illustrates the problem wherein a person on the left side of the figure is running while a person on the right is jogging. The traditional

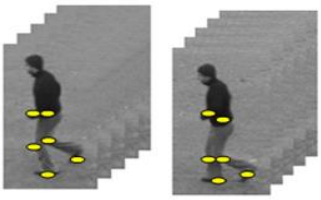


Fig.1. Similar activities (a person on the left side is running; a person on the right side is jogging)

Bag-of-words based methodology often treats them as the same action which is wrong. To overcome this limitation researchers have used the spatiotemporal contextual information [32, 33, 34]. But the limitation of these methods is that they are still weak in noise tolerance. When the noise is added, the representation of the interest points of action might be recognized as other activities. Moreover, the measures of contextual distance between spatiotemporal points traditionally use Euclidean distances among these points but this distance cannot hold the structural information of clusters. The proposed methodology is capable to overcome these limitations.

III. Proposed Methodology

Local features such as spatiotemporal points are most popularly used to represent an action. Spatiotemporal interest points such as Harris 3D detectors and SIFT detectors [43] are used to detect the interest points. We used the Harris 3D detector to extract the spatio-temporal interest points. These interest points are described locally by the descriptors such as HOG [27] and HOF [28]. The HOG and HOF are used in the proposed method to describe the interest points detected by Harris 3D detector. These local descriptors represent the feature vectors of each frame of the action video. A bag-of-visual-word of spatiotemporal interest points is formed to represent an action efficiently. The codebook is created by performing vector quantization through clustering. The K-means clustering [32] is popularly used. In K-means clustering clusters are placed near to most frequently occurring feature vectors. These most frequently occurring feature vectors are used as the discriminative features of action. But for certain actions, these most frequently occurring features are not capable enough to show the discriminative nature of an action. Further quantization error may also occur due to the assignment of a single cluster to feature vector. To avoid this we used the soft assigning technique [15] where Gaussian kernel is used in which feature vectors and cluster are represented by the normal distribution. Equation (1) defines the visual word $V_{i,t}$

$$V_{i,t} = \exp^{\left(-\frac{\|f_t - C_i\|^2}{2\theta^2}\right)} \quad (1)$$

where, f_t is the feature vector and C_i is the i^{th} cluster where $i=1, 2, 3\ldots N$ and θ is the smoothing parameter. We used the cross-validation to find out the size of the Gaussian kernel. In the proposed method we used this soft assigning technique and we get the codebook $\{C_1, C_2 \ldots C_N\}$ where C_i is the i^{th} cluster and the total number of the cluster is N. Fig. 2. shows the proposed methodology for codebook creation. Let the spatiotemporal feature points be represented in the form of tuple $X = (L, r)$ where L is local descriptor (appearance) of the interest points and $r = (x, y, t)$ is the spatial location of the interest points. The horizontal, vertical and temporal coordinates of interest points are represented by x, y, and t respectively. As discussed above, in the traditional bag-of-visual-word based methodology, the geometrical structure of the cluster cannot be maintained because of the introduction of noise. This results in the reduction of the recognition rate. Thus, to maintain the geometrical structure of the clusters, we propose a new methodology where we calculate the contextual distance among the points of the cluster on the basis of the difference between the contributions of points in a cluster to maintain its geometrical structure. The Methodology can be understood by the block diagram in fig. 2.

The spatiotemporal points are detected from the input action video by using Harris 3D corner detector. These points are described by the descriptors HOG and HOF. Using soft assigning technique we created the codebook having N no. of clusters. The traditional bag-of-visual-word technique is modified through the proposed methodology. Using the geometrical descriptor, the contextual distance among the points of the cluster is calculated. The directed graph of the cluster points is then created with the help of contextual distance between the points. These directed graphs are described by Laplacian.

A. Geometrical Description of a Cluster

As discussed above, among the N clusters formed, let us select the i[th] cluster C_i. The cluster C_i is represented by the set I_c. The set I_c is represented by (2) where i_c is a mean point of the cluster C_i having m nearest neighboring points.

$$\mathrm{I}_c = \left(i_{c0}, i_{c1}, i_{c2}, ..., i_{cm}\right) \quad (2)$$

We used centroid as the geometrical structure descriptor of set I_c denoted by (3) as shown below

$$\mathrm{M}\left(\mathrm{I}_c\right) = \frac{1}{m+1}\sum_{j=0}^{m} i_{ck} \,. \quad (3)$$

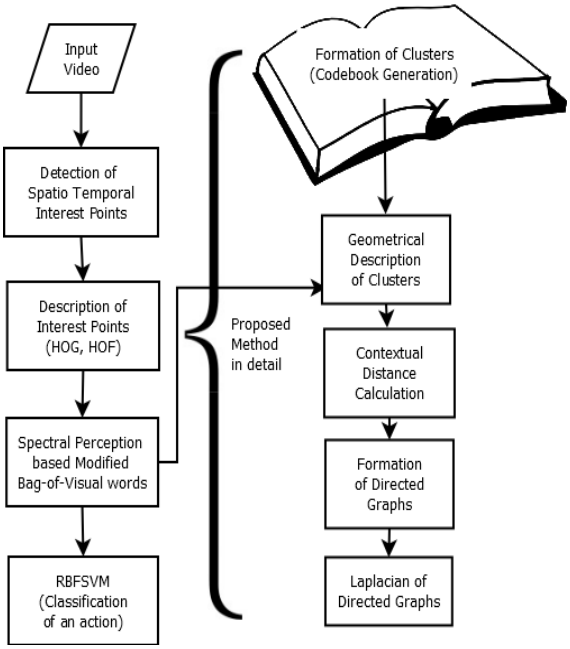


Fig.2. Detail process flow diagram of the proposed methodology

To calculate the contribution of each point in the set I_c, we removed one point at a time and recalculated the centroid using (3). Thus we get the contribution of all points in the set I_c that forms the geometrical structure of the cluster. The contribution of point i_{ck} can be found out by (4).

$$\delta\left(i_{ck}\right) = \left|\mathrm{M}\left(\mathrm{I}_c\right) - \mathrm{M}\left(\mathrm{I}_c \big/ i_{ck}\right)\right| \quad (4)$$

where $\delta(i_{ck})$ is the contribution of the point i_{ck} in the cluster set I_c. $M(I_c)$ is the complete descriptor of I_c and $M(I_c/i_{ck})$ is the description of set I_c after removing the point i_{ck} from the set. In a similar manner contribution of all the points in a cluster's set is found out. If the value of $\delta(i_{ck})$ is very small then the contribution of the point i_{ck} is not significant in maintaining the geometrical structure of the cluster and if the value of $\delta(i_{ck})$ is high then we can conclude that the point i_{ck} has a significant effect in retaining the geometrical structure of the set I_c.

B. Contextual Distance Among the Points of a Cluster

Fig. 3 describes the geometrical description of the cluster and the contextual distance among the points of the clusters. Once the contribution of points of a cluster to maintain its geometrical structure has been obtained, the contextual distance among these points is calculated by taking the difference between their contributions. This can be calculated from (5).

$$\Psi\left(i_{c0}, i_{c1}\right) = \left\{\delta\left(i_{c0}\right) - \delta\left(i_{c1}\right)\right\} \quad (5)$$

where $\Psi(i_{c0},i_{c1})$ is the contextual distance between interest points i_{c0} to i_{c1} and $\delta(i_{c0})$ *and* $\delta(i_{c1})$ are the contribution of the points i_{c0} and i_{c1}. After calculating the contextual distances among the points of the cluster set I_c the set I_c is represented in the form of their contextual distances. This representation of the cluster set creates asymmetry among the points of the set I_c because the contextual distance between i_{c0} to i_{c1} may not be equal to the contextual distance between i_{c1} to i_{c0}. To deal with this naturally occurred asymmetry in the cluster set I_c, the directed graph has been used to make the cluster more informative as it introduced direction also.

C. Formation of the Directed Graph of a Cluster on the Basis of Contextual Distance

The cluster set I_c can be represented by the directed graph where the points of set are treated as the nodes/vertices of the graph as shown in fig. 4. Let the edge from the points i_{c0} to i_{c1} is defined by the weight $\lambda(i_{c0},i_{c1})$ as given by (6).

$$\lambda\left(i_{c0},i_{c1}\right)=e^{\left[\frac{\Psi\left(i_{c0},i_{c1}\right)}{\mu}\right]^2} \tag{6}$$

where $\lambda(i_{c0},i_{c1})$ is the weight of the edge from point i_{c0} to i_{c1}, $\Psi(i_{c0},i_{c1})$ is the contextual distance from i_{c0} to i_{c1}, μ is a free parameter and i_{c0} and i_{c1} are the neighborhood points of the cluster set I_c. Thus, in the cluster set I_c a point i_c is connected to its m neighborhood points through m edges as shown in fig. 4.

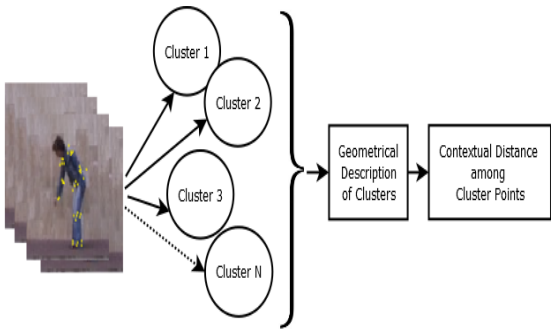


Fig.3. Geometrical description of a cluster and contextual distance among the cluster points

The directed graph of cluster set can be represented through weighted edge matrix W. The weighted edge matrix W carries the geometrical structural information of the cluster set I_c. Similarly, directed graphs have been applied on other contextual cluster sets also. The weighted edge matrix is created for a single cluster as shown below:

$$\begin{bmatrix} & i_{c0} & i_{c1} & i_{c2} & \cdots & i_{cm} \\ i_{c0} & \lambda(i_{c0},i_{c0}) & \lambda(i_{c0},i_{c1}) & \lambda(i_{c0},i_{c2}) & \dots & \lambda(i_{c0},i_{cm}) \\ i_{c1} & \lambda(i_{c1},i_{c0}) & \lambda(i_{c1},i_{c1}) & \lambda(i_{c1},i_{c2}) & \dots & \lambda(i_{c1},i_{cm}) \\ i_{c2} & \lambda(i_{c2},i_{c0}) & \lambda(i_{c2},i_{c1}) & \lambda(i_{c2},i_{c2}) & \dots & \lambda(i_{c2},i_{cm}) \\ \vdots & \dots & \dots & \dots & \ddots & \vdots \\ i_{cm} & \lambda(i_{cm},i_{c0}) & \lambda(i_{cm},i_{c1}) & \lambda(i_{cm},i_{c2}) & \cdots & \lambda(i_{cm},i_{cm}) \end{bmatrix}$$

Further to describe the weighted directed graph Laplacian of the directed graph has been used.

D. Laplacian of the Directed Graph of a Cluster

Suppose $G(i_{cj},\lambda)$, where $j=0\dots m$, is the directed graph of the set I_c and the directed edge (i_c, i_d) is the edge directed from vertex i_c to vertex i_d. A walk from one vertex to other is represented by the transition probability matrix O. If $O(i_c, i_d)$ is the transition probability of moving from point i_c to i_d, it is obvious that $O(i_c, i_d)>0$ only if there is an edge between i_c to i_d. Further for the weighted directed graph, the transition probability matrix is given by (7).

$$\mathrm{O}\left(i_c,i_d\right)=\frac{\lambda_{(i_c,i_d)}}{\sum_m \lambda_{(i_c,i_m)}} \tag{7}$$

wherein, equation $\lambda(i_c, i_d)$ is the weight of the edge from point i_c to i_d and denominator in the right-hand side of the equation represents the total weighted sum of all the edges from point i_c to all its neighboring points.

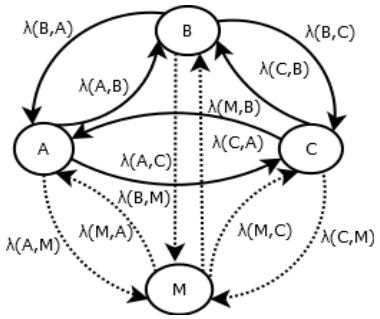


Fig.4. A cluster set I_c of m points where A=i_{c0}, B=i_{c1} and λ (A, B) = (i_{c0},i_{c1})

According to (7) transition probability matrix of the directed graph has a unique left Eigenvectors ε with all positive values and the Laplacian of the transition probability matrix Φ is defined by the (8).

$$\Phi=\mathrm{I}-\frac{\epsilon^{\frac{1}{2}}\,O\,\epsilon^{-\frac{1}{2}}+\epsilon^{-\frac{1}{2}}\,O\,\epsilon^{\frac{1}{2}}}{2} \tag{8}$$

where O is the transition probability matrix and ε is the Eigenvector. Equation (8) gives the Laplacian of the transition probability matrix for a cluster C_i. This Laplacian of the transition probability matrix is converted to a vector having dimension by scanning the matrix from the top-left element to bottom right element. Thus, the Laplacian of the transition probability matrix for N clusters has N feature vectors. These feature vectors are fed to RBF-SVM classifier. Fig. 5 shows the formation of N clusters and their Laplacian descriptors.

IV. Experimental Results

The proposed method has been implemented in MATLAB R2013a having hardware configuration processor Intel(R) Core (TM) i5-6200U CPU @2.30GHz 2.40 GHz with 12 GB RAM and 64-bit operating system. For evaluation of the performance of proposed methodology two evaluation factors, accuracy and the equal error rate is used. Accuracy gives the correct prediction of the action out of all the predictions. Three standard datasets KTH [35], IXMAS[36] and Ballet[20] have been used to validate the proposed methodology. IXMAS dataset has been used for parameter settings and the same setting is applied to other datasets. For cross-validation, leave one actor out strategy has been adopted.

A. Parameter Settings

There are two parameters that can affect the proposed methodology. The First parameter is the size of the codebook using the soft assigning strategy and second is the number of nearest neighbors of a cluster. In most of the bag-of-visual-word based methods, the size of the codebook is around 1000. But in the proposed method cross-validation optimization has been done to find out the size of codebook. Fig. 6.a shows the codebook size 25, 30, 35 40, 45 and 50. The codebook size of 35 has been opted, because if the size of the cluster is greater than 30 then the accuracy varies only 2-3%. Another parameter is the number of nearest neighbor in the cluster. Different values such as 100, 120, 140, 160, 180, 200 and 220 have

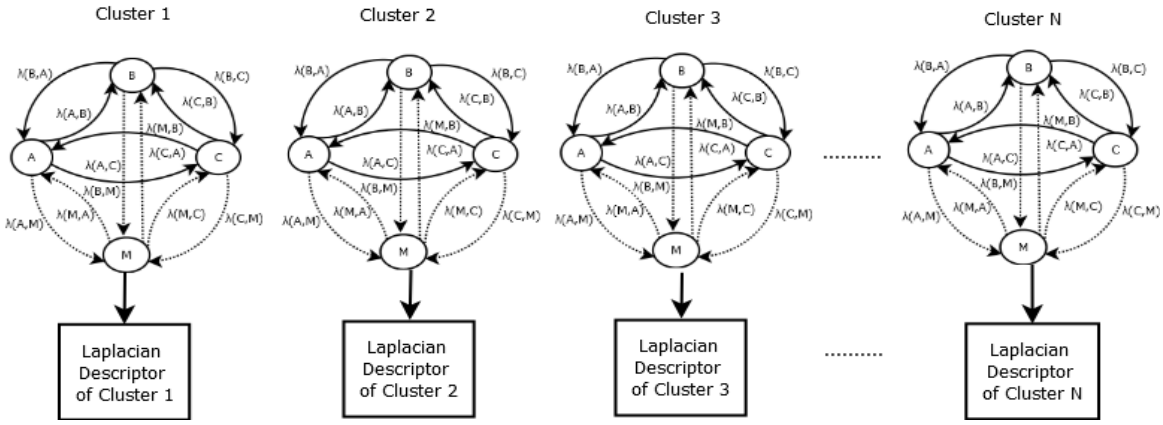
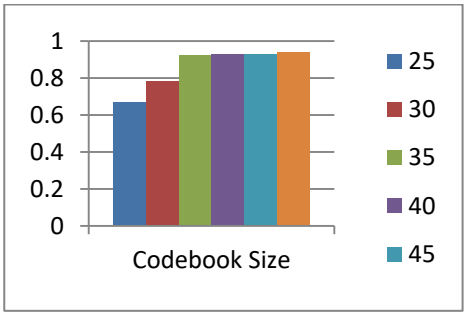
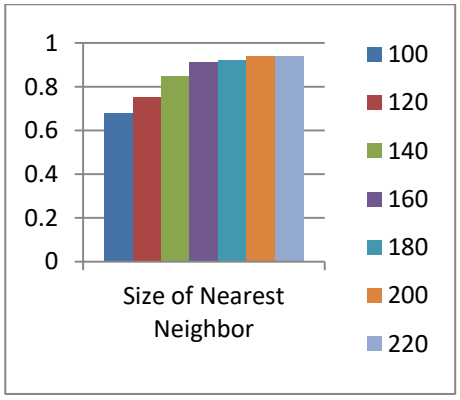


Fig.5. Formation of N clusters and their Laplacian descriptors



(a)



(b)

Fig.6. (a). Selection of codebook size (b) Selection of nearest neighbor size

been taken. The size of nearest neighbor 160 has been taken because other than this value the accuracy varies only 1-2% as shown in fig. 6.b. Similar setting is used for KTH and Ballet datasets.

B. Analysis and Discussion on KTH Dataset

The KTH dataset [35] comprises of six actions such as hand-applauding, waving, boxing, walking, jogging and running shown in fig. 7.a. Every activity has 100 recordings for four unique situations in various light conditions, indoor and open-air conditions.

Fig. 8 shows the confusion matrix where the proposed methodology gets the appreciable result for differentiating different actions. Similar types of actions like jogging and running are also very well classified. Only 5% jogging action is wrongly classified as running and running is only 3% confused with the jogging. The proposed method classifies hand waving action accurately, but it gets 7% confused with the action waving while classifying the boxing action.

The proposed methodology has been compared with the other state-of-the-art methods for KTH dataset in Table 1. Sadek et.al [10] uses the affine moment invariants of 3D action volume where they get the average accuracy of 93.3%. But to create the 3D volume of action there is a requirement of a sophisticated approach of background subtraction to detect the object.

The action is represented with the combination of pose descriptors and negative space descriptor of each frame [16]. They get higher accuracy of 94.4% but still, there is also required to have a very good background subtraction technique. Furthermore, [18] and [19] are getting very high accuracies of 95.8% and 96% respectively wherein [18] proposed a new feature Hanklets which are view-invariant features and makes the accuracy higher and [19] is traditional deep learning-based methodology for action recognition that is presently used more popularly. The proposed methodology is outperforming these methodologies on KTH dataset as it is robust to view change, illumination change, and noise also. Table 1 shows that 98.2% of action samples are correctly classified which is a very high accuracy as compared to other prominent methods of Table 1.

Table 1. Comparison among other state-of-the-art methods for KTH dataset

Method	Year	EER	Accuracy
[30]	2015	5.8	94.2
[31]	2015	6.8	93.1
[10]	2012	6.7	93.3
[15]	2013	7.4	92.6
[16]	2014	5.6	94.4
[17]	2015	8.7	91.3
[18]	2012	4.2	95.8
[19]	2017	3.2	96.8
Proposed Method		**1.8**	**98.2**

In Table 2 comparison among the similar state-of-the-art methods is shown. The methodologies [32, 37-41] are based on the spatiotemporal features interest points. The disadvantage of these methodologies is that they do not show the relationship between the interest points and also the structural information of the clusters is missing. Although [32] retains the relationship among the spatiotemporal interest point, it is immune to noise. The proposed method shows the superior accuracy of 98.2% because it deals with these issues efficiently as discussed in the previous sections.

C. Analysis and Discussion on Ballet Dataset

The Second dataset that has been used is Ballet dance [20] shown in fig. 7.b. It is a typical dance dataset where different expressive dance steps. A single artist is performing the dance act. The dataset contains distinctive acts in Ballet dance such as hopping, jumping, standing still, left to right-hand opening, right to left-hand opening, leg swinging and turning right. This dataset has a total of 44 numbers of videos.

Proposed methodology maintains the structural information of the clusters and thus classifies similar actions better than other methods as summarized in Table 3. Fig. 9 shows the confusion matrix for actions in Ballet dataset. It is clear from the matrix that actions such as standing still, leg swinging and turning right are accurately classified by the proposed method. Whereas hopping is 5% confused with jumping and jumping action is 8% confused with hopping. Similarly, Left to right-hand opening is 10% confused with right-hand opening and right to left-hand opening is 12% confused with the left to right-hand opening. Table 3 shows the comparison of the proposed method with some very efficient methods when they tested on Ballet datasets. Vishwakarma et.al [5] uses the silhouettes as the key poses that can express the actions. These silhouettes are represented into grids and cells. A fusion of classifier is used to recognize the actions. They achieved an accuracy of 94.2%. Their only disadvantage is that to represent poses the accurate silhouettes are required. Methods [22] and [32] represented the action with the set of spatiotemporal descriptors. Where [22] uses the dictionary learning approach and [32] uses the contextual constraint linear coding for action recognition. They showed 94.2% and 91.2% accuracy respectively. These methods are not robust to the data tempered by the noise. But as the proposed method is robust to noise that makes its accuracy comparable to these methods. The accuracy and the equal error rate of the proposed methodology are compared with other promising state-of-the-art methodologies in Table 3. The proposed methodology achieved 96.2% accuracy and 3.8% equal error rate.

Table 2. Comparison among similar state-of-the-art methods for KTH dataset

Method	Year	EER	Accuracy
[37]	2014	5.8	95.0
[38]	2005	6.8	76.4
[39]	2011	6.7	84.1
[40]	2012	7.4	93.6
[41]	2013	7.5	92.5
[32]	2012	5.7	94.3
Proposed Method		**1.8**	**98.2**

D. Analysis and Discussion on IXMAS dataset

The third dataset that the proposed methodology experimented on is IXMAS. There are five different cameras used for the videos in IXMAS dataset. For IXMAS dataset [36] five camera views are taken. The proposed methodology has been compared with some latest prominent methodologies in Table 4. In Table 3 the methodologies [23] and [26] achieved the average accuracy of 87.6% and 87.1% on IXMAS dataset. [23] is based on the multi-view human action recognition where several cameras are observing a scene and features from each camera are considered for recognition. It makes the system complex. In [26] a statistical translation framework is used where the cross-view approach is applied. The proposed methodology achieved 90% accuracy for all five camera views. In the proposed method the camera view 3 shows the highest accuracy of 92.4%. The overall accuracy computed by taking an average of all camera views is 90.5%.

Table 3. Illustration of EER and Accuracy for various approaches on Ballet dataset

Method	Year	EER	Accuracy
[20]	2008	48.7	51.3
[27]	2009	8.7	91.3
[21]	2014	9.2	90.8
[28]	2014	8.8	91.2
[32]	2012	5.5	94.5
[5]	2015	5.8	94.2
[22]	2012	8.8	91.2
Proposed Method		**3.8**	**96.2**

Table 4. Accuracy for various approaches from 5 different cameras

Method	[23]	[29]	[24]	[25]	[26]	**Proposed Method**
Year	2011	2010	2012	2013	2016	
C1	89.1	84.2	86.5	91.3	88.4	**90.8**
C2	83.4	85.2	83.8	85.7	85.3	**90.6**
C3	89.3	84.1	86.1	89.3	88.3	**92.4**
C4	87.2	81.5	84.5	90.2	86.5	**88.5**
C5	89.2	82.6	87.4	86.5	87.2	**90.6**
Overall Accuracy	87.6	83.5	85.6	88.6	87.1	**90.5**

(a)

(b)

Fig.7. (a). KTH dataset; (b). Ballet dataset

	Applauding	**Waving**	**Boxing**	**Walking**	**Jogging**	**Running**
Applauding	**1.00**	0.00	0.00	0.00	0.00	0.00
Waving	0.00	**1.00**	0.00	0.00	0.00	0.00
Boxing	0.00	0.07	**0.93**	0.00	0.00	0.00
Walking	0.00	0.00	0.00	**1.00**	0.00	0.00
Jogging	0.00	0.00	0.00	0.00	**0.95**	0.05
Running	0.00	0.00	0.00	0.00	0.03	**0.97**

Fig.8. Confusion Matrix for actions in KTH dataset

	Hopping	**Jumping**	**Standing Still**	**LR Hand Opening**	**RL Hand Opening**	**Leg Swinging**	**Turning Right**
Hopping	**0.95**	0.05	0.00	0.00	0.00	0.00	0.00
Jumping	0.08	**0.92**	0.00	0.00	0.00	0.00	0.00
Standing Still	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00
LR Hand Opening	0.00	0.00	0.00	**0.90**	0.10	0.00	0.00
RL Hand Opening	0.00	0.00	0.00	0.12	**0.88**	0.00	0.00
Leg Swinging	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00
Turning Right	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**

Fig.9. Confusion Matrix for actions in Ballet dataset

	Check Watch	**Cross Arms**	**Scratch Head**	**Get Up**	**Turn Around**	**Walk**	**Wave**	**Punch**	**Kick**	**Pick Up**
Check Watch	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cross Arms	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Scratch Head	0.00	0.00	**0.95**	0.00	0.00	0.00	0.05	0.00	0.00	0.00
Get Up	0.00	0.00	0.00	**0.93**	0.00	0.00	0.00	0.00	0.00	0.07
Turn Around	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00
Walk	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00
Wave	0.00	0.00	0.08	0.00	0.00	0.00	**0.92**	0.00	0.00	0.00
Punch	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
Kick	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00
Pick Up	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	**0.90**

Fig.10. Confusion Matrix for actions in IXMAS dataset

There are different challenges included in this dataset that can reduce the recognition such as appearance change due to change in viewpoint and numbers of similar types of activities. Fig. 10 depicts the overall confusion matrix where the action scratching head is only 5% confused with the action wave. Wave is 8% confused with the action scratching head. Getting up is 7% confused with the action picking up and picking up is 10% confused with the getting up. The confusion percentage is very less for these similar activities shows the dominance of the proposed methodology in the case of interclass similarity. Other actions are clearly classified with 100% classification.

To summarize, the overall performance of the proposed method shows that this is a very effective method for human action recognition. The three action datasets comprise of several challenges for human action recognition. The proposed method achieved 98.2% accuracy for KTH dataset, 96.2% accuracy for Ballet dataset and the average accuracy for all five camera views of the IXMAS datasets is 90.5%.

V. Conclusion

In this research paper, a new methodology that is the modified version of the bag-of-visual-word has been proposed. The limitation of the traditional codebook based method where the geometrical structural information of the clusters was missing has been overcome in the proposed methodology. Moreover, a new methodology for calculating the contextual distance among the points of the cluster has been used for action recognition. The proposed methodology has been validated on the challenging public dataset such as KTH, Ballet, and IXMAS where it shows its superiority as compared to other existing methods.

Acknowledgment

The authors wish to thank the Department of Electronics and Communication Engineering of Delhi Technological University, Delhi India for their support.

References

[1] J. Aggrawal, and M. Rayoo, “Human Activity Analysis: A Review,” *Journal of ACM Computing Survey*, vol. 43(3), pp. 16-43, 2011.

[2] J. Dou, and J. Li, "Robust human action recognition based on spatiotemporal descriptors and motion temporal templates," *Optik*, vol.125(7), pp. 1891-1896, 2014.

[3] D.K. Vishwakarma, R. Kapoor, and A. Dhiman, “A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics,” *Journal of robotics and autonomous system*, vol. 77, pp. 25-28, 2016.

[4] D. Zhao, L. Shao, X. Zhen, and Y. Liu, “Combining appearance and structural features for human action recognition,” *Neurocomputing*, vol. 113(3), pp. 88-96, 2013.

[5] D.K. Vishwakarma, and R. Kapoor, “Hybrid classifier based human activity recognition using the silhouette and cells,” *Expert Systems with Applications*, vol. 42(20), pp. 6957-6965, 2015.

[6] C. Achar, X. Qu, A. Mokhber, and M. Milgram, “A novel approach for action recognition of human actions with semi-global features,” *Journal of machine vision and application*, vol. 19 (1), pp.27-34, 2008.

[7] C. Lin, F. Hsu, and W. Lin, “Recognizing human actions using NWFE-based histogram vectors,” *EURASIP Journal of Advances in Signal Processing*, vol. 9, 2010.

[8] J. Gu, X. Ding, and S. Wang, “Action recognition from arbitrary views using 3D-key-pose set,”*Frontiers of Electrical and Electronic Engineering*, vol. 7(2), pp.224-241, 2012.

[9] L.Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, “Actions as Space-Time Shapes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29(12), pp.2247-2253, 2007.

[10] S. Sadek, A. Hamadi, M. Elmezain, B. Michaelis, and U. Sayed, “Human action recognition via affine moment invariants,” *International Conference on Pattern Recognition, Tsukuba*, pp. 218-221,2012.

[11] L. Presti, M. Cascia, S. Sclaroff, and O. Camps, “Hanklet-based dynamical systems modeling for 3-D action recognition,” *Journal Image and Vision Computing*, vol. 44(C), pp.29-43, 2015.

[12] A. Yilmaz, and M. Shah, “A differential geometric approach to representing the human actions,” *Computer Vision and Image Understanding (CVIU)*, pp .335-351 2008.

[13] K. Guo, P. Ishwar, and J. Konrad, “Action recognition from video using feature covariance matrices,” *IEEE Transaction on Image Processing*, vol. 22(6), pp. 2479-2494, 2013.

[14] Q. Li, H. Cheng, Y. Zhou, and G. Huo, “Human Action Recognition Using Improved Salient Dense Trajectories,” *Computational Intelligence and Neuroscience*, vol.5,pp.1-11, 2016.

[15] D. Wu, and L. Shao, “Silhouette analysis-based action recognition via exploiting human poses,” *IEEE transaction on circuits and systems for video technology*, vol. 23(2), pp. 236-243, 2013.

[16] S. A. Rahman, I. Song, M.K.H. Leung, I. Lee, and K. Lee, “Fast action recognition using negative space features,” *Expert System and Applications*, vol. 41(2), pp.574-587, 2014.

[17] I.G. Conde, and D.N Olivieri, “A KPCA spatiotemporal differential geometric trajectory cloud classifier for recognizing human actions in a CBVR system,” *Expert System and Applications*, vol. 42(13), pp.5472-5490,2015.

[18] B. Li, O.I. Camps, and M. Sznaier, “Cross-view activity recognition using Hanklets,” *International Conference on Computer Vision and Pattern Recognition, Providence*, 2012.

[19] Y. Shi, Y. Tian, Y. Wang, and T. Huang, “Sequential deep trajectory descriptor for action recognition with three-stream CNN,” *IEEE Transaction of Multimedia*, vol. 19(7), pp.1510-1520, 2017.

[20] A. Fathi, and G. Mori, “Action recognition by learning mid-level motion features,” *International Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage*, 2008.

[21] X.L. Min, H.J. Xia, and T.L. Zheng, “Human action recognition based on chaotic invariants,”*Journal of South Central University*, pp. 3171-3179, 2014.

[22] Guha, T. and Ward, R.K., “Learning sparse representations for human action recognition,” IEEE

Transaction of Pattern Analysis and Machine Intelligence, 34(8), pp. 1576-1588, 2012.
[23] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using con-text and appearance distribution features," *International IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence*, 2011.
[24] X. Wu and Y. Jia, "View-invariant action recognition using latent kernelized structural SVM," *Proceedings of the 12th European Conference on Computer Vision (ECCV), Florence, Italy*, 2012.
[25] E.A. Mosabbeb, K. Raahemifar, and M. Fathy, "Multi-view human activity recognition in distributed camera," *Sensors*, vol. 13(7), pp. 8750-8770, 2013.
[26] J. Wang, H. Zheng, J. Gao, and J. Cen, "Cross-view action recognition based on a statistical translation framework," *IEEE Transaction of Circuits Systems for Video Technology*, vol. 26(8), pp.1461-1475, 2016.
[27] Y. Wang, and G. Mori, "Human action recognition using semi-latent topic model," *IEEE Transactional of Pattern Analysis and Machine Intelligence*, vol. 31(10), pp. 1762-1764, 2009.
[28] A. Iosifidis, A. Tefas, and I. Pitas, "Discriminant bag of words based representation for human action recognition," *Pattern Recognition Letters*, vol. 49(C), pp. 185-192, 2014.
[29] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," *European Conference on Computer Vision (ECCV), Crete, Greece*, 2010.
[30] L. Wang, Y. Qiao and X. Tang, "Action recognition with trajectory pooled deep-convolutional descriptors," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4305-4314, 2015.
[31] L. Sun, K. Jia, D-Y. Yeung, and B.E. Shi, "Human action recognition using factorized spatiotemporal convolutional networks," *IEEE International Conference on Computer Vision (ICCV)*, pp. 4597-4605, 2015.
[32] Z. Zhang, W. Chunheng, X. Baihua, Z. Wen, and L. Shuang, "Action Recognition Using Context-Constrained Linear Coding," *IEEE Signal Processing Letters*, vol. 19(7), pp. 439-442, 2012.
[33] A. Kovashka, and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2046-2053, 2010.
[34] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfield, "Learning realistic human actions from movies," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1-8, Jun.2008.
[35] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," *Proceedings of the 17th International Conference on Pattern Recognition, Cambridge*, 2004.
[36] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104(2-3), pp. 249-257, 2006.
[37] H. Liu, M. Liu, and Q. Sun, "Learning directional co-occurrence for human action classification," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1235-1239, 2014.
[38] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatiotemporal features," *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65-72, 2005.
[39] M. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," *IEEE International Conference on Computer Vision (ICCV)*, pp. 1036-1043, 2011.
[40] Q. Sun, and H. Liu, "Action disambiguation analysis using normalized Google-like distance correlogram," *Springer Asian Conference on Computer Vision (ACCV)*, pp. 425-437, 2012.
[41] Q. Sun, and H. Liu, "Learning spatiotemporal co-occurrence correlograms for efficient human action classification," *IEEE International Conference on Image Processing (ICIP)*, pp. 3220-3224, 2013.
[42] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality constrained linear coding for image classification," *Proc. CVPR*, pp. 3360-3367, Jun. 2010.
[43] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64(2-3), pp. 107-123, Sep.2005.

Authors' Profiles

Mr. Om Mishra is a research scholar in Department of Electronics & Communication, Delhi Technological University, Delhi, India. He has worked as an Assistant Professor in G.B. Pant Government Engineering College, New Delhi, India. He received Master of Engineering in Electronics & Communication Engineering from Delhi College of Engineering (Presently DTU), Delhi, India. His Research interest includes Vision based Activity Recognition, Signal Processing, Pattern Recognition.

Dr. Rajiv Kapoor is a Professor in Delhi Technological University, Delhi, India. He worked as Principal in AIACTR, Delhi, India in diverted capacity. He is Ph.D. in Electronics and Communication Engineering from Punjab Engineering College, India. His Research interest include Vision/Speech based Tracking, Activity Recognition Vision/Speech based, Signal Processing, Pattern Recognition. He has published more than 100 research papers in International Journals and Conferences. He is also an author of books published in IGI Global and Springer. He has been a Principal Investigator of sponsored projects, Government of India.

Dr. M.M. Tripathi is a Professor in Electrical Engineering Department of Delhi Technological University, Delhi, India. He has also worked as Scientist with the Institute for Plasma Research, India and National Institute of Electronics & Information Technology, India. He is Ph.D. in Electrical Engineering from G. B. Technical University, India. His research interests include Artificial Intelligence applications, Renewable energy and Power system

restructuring. He has published more than 30 research papers in International Journal and Conferences. Presently, he is Director, IQAC, Delhi Technological University, Delhi.

How to cite this paper: Om Mishra, Rajiv Kapoor, M.M. Tripathi, "Human Action Recognition Using Modified Bag of Visual Word based on Spectral Perception", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.11, No.9, pp. 34-43, 2019.DOI: 10.5815/ijigsp.2019.09.04

