International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-10 Issue-4, April 2021

Extracting Multiple Features for Dynamic Hand Gesture Recognition

Suni S. S, K. Gopakumar

Abstract*: In this work a framework based on histogram of orientation of optical flow (HOOF) and local binary pattern from three orthogonal planes (LBP_TOP) is proposed for recognizing dynamic hand gestures. HOOF algorithm extracts local shape and dynamic motion information of gestures from image sequences and local descriptor LBP is extended to three orthogonal planes to create an efficient motion descriptor. These features are invariant to scale, translation, illumination and direction of motion. The performance of the new framework is tested in two different ways. The first one is by fusing the global and local features as one descriptor and the other is using features separately to train the multi class support vector machine. Performance analysis shows that the proposed approach produces better results for recognizing dynamic hand gestures when compared with state of the art methods.

Keywords*: Hand gesture recognition, Histogram of orientation of optical flow, local binary pattern, Multiclass support vector machine, Scale invariant feature.

I. INTRODUCTION

The unique nature of human appearance, variety of apparels and background clutter in video sequences makes detection of human gestures next to impossible. To make human computer interaction seamless as possible, a huge amount of research and development efforts are going on [1, 2]. Hand gestures are widely used as a natural human computer interface that forces the researchers to work more on gesture characteristics, its representations and detection techniques [3,4]. Still, certain human movements are very characteristics, strictly individual specific, so detector performance can potentially be reformed by including motion information. Most of the detector tool works by keeping the camera and background static. Overall performance of the system is greatly enhanced by this. The developmental challenge lies in its ability to select a set of local and global parameters that can analyze the human motion effectively.

Recognition of hand gestures is based on the three main stages i.e. hand segmentation, tracking and identification. This paper proposes the framework of combining two features; histogram of orientation of optical flow (HOOF) and local binary pattern from three orthogonal planes (LBP_TOP) to develop the motion descriptor for categorizing the various hand gestures. Multiclass support vector machine classifier is utilized to recognize the hand gestures using the multiple feature based motion descriptor. The rest of the article is organized as follows. Related works are detailed in Section 2. Section 3 deals with proposed architecture. Section 4 deals with the experimental results and its analysis. Finally, conclusion and future scope are given in Section 5.

II. RELATED WORKS

Hand gesture recognition involves the following stages; pre-processing and hand segmentation, feature extraction, its representation and recognition. Pre-processing techniques like filtering and edge detection are done to reduce the noise and the hand region is segmented from the image frames. This segmented hand image is used to extract the feature vectors that can represent the particular gesture. These discriminating features are used to create the models. Classifiers are trained with these feature models and then identify the new feature models. Global and local approaches of extracting features are used for representing hand gestures [5,6].

Globally, optical flow measures the displacement between two consecutive frames. Optical flow refers to the study of object motions in foreground under static background. In [7], histograms based on optical flow were utilized to match the movement of a soccer player to that of object in a control video. Different motion descriptors based on optical flow and its variants which uses silhouette of a person are presented in Tran et al [8]. Most of the methods focused on local representation of features not on temporal dynamics characteristics. In this work, Histogram of Oriented Optical Flow (HOOF) algorithm is applied to extract the features from hand gesture images. This global feature is invariant to scale and direction of movement. One of the major attractions of HOOF features is that it does not require any hand detection and background elimination process.

Mostly, local approaches use spatiotemporal features for representing motion in a video [9,10]. Niebles [9] proposed an unsupervised learning technique for probability distributions of spatial key points in human action videos. Zhao et. al. [10] introduced dynamic texture based features such as local binary pattern from three orthogonal planes (LBP – TOP) and volumetric local binary pattern (VLBP) for the recognition of facial expressions. But the number of pattern of basic VLBP is increases as the number of neighboring points increases, which restricts the usages in various applications.

Manuscript received on March 22, 2021.
Revised Manuscript received on April 05, 2021.
Manuscript published on April 30, 2021.
* Correspondence Author
Suni S S*, Research Scholar, LBS Centre for Science and Technology, University of Kerala, Thiruvananthapuram, Kerala, India.
K Gopakumar, Professor, TKM College of Engineering, Kollam, Kerala, India

In this work, the dynamic textures are extracted using local binary patterns from three orthogonal planes. These computationally simple features are spatiotemporal and are invariant to different illuminations.

The main limitations in gesture recognition process are tracking the hand in presence of noise, self-occlusions, and recognition of gestures in changes of camera viewpoints, variations in pose and scale.

In this work, a framework is created by fusing both HOOF with LBP-TOP features. It extracts the appearance and motion information to create an efficient feature descriptor. It is proved that the combined or fused feature descriptor gives better performance.

III. PROPOSED SYSTEM OVERVIEW

This paper concentrates on developing an architecture that effectively fuses global and local features to build motion descriptor for recognizing the hand gestures. Fig.1. shows the framework for hand gesture recognition.

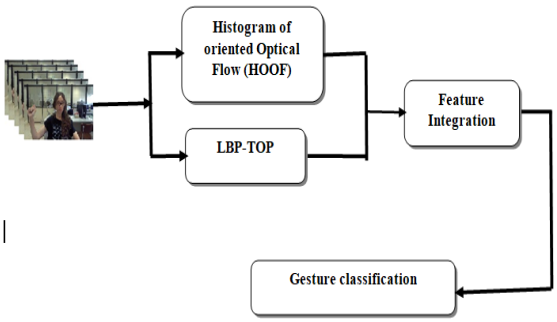

Fig. 1. Proposed framework for hand gesture recognition

A. Histogram of Oriented Optical Flow (HOOF)

Inspired by the developments and its success of histogram of gradients and its variants of features in the object detection group, we postulate that the simple and efficient feature to use in movement detection is optical flow. Unfortunately, the pixel dynamics changes over time as well liable to background variations like scale changes, noise and direction to arrive at a conclusive data. To avoid these problems, distribution flow of motion is set as the standard since the motion of the hand under stationary background induces similar characteristic profile. For example, a zoomed hand gesture versus a normal gesture of a person, the optical flow vector magnitude will be larger in the case of zoomed hand gesture. Same is the case when the movement of hand is from the left to the right, as it will be reflected in the vertical axis to that of the movement in the opposite direction. Thus the need for a gesture profile based on optical flow at a given time is invariant to directionality of movement and the scale. To overcome such challenges, we have proposed the Histogram of Oriented Optical Flow (HOOF) algorithm to extract features from the images. Following procedure is used to evaluate HOOF feature vector. Initially, the optical flow is calculated for every frame of the videos. According to the primary angle from the horizontal direction every flow vector is binned and given weight corresponding to the magnitude.

Hence the flow vectors, $\vartheta = [x, y]^T$ with direction, $\theta = tan^{-1}\left(\frac{y}{x}\right)$ in the range

$$-\frac{\pi}{2} + \pi\frac{b-1}{N} \leq \theta < -\frac{\pi}{2} + \pi\frac{b}{N} \qquad (1)$$

will contribute by $\sqrt{x^2 + y^2}$ to the sum in bin b, $1 \leq b \leq N$ out of total of N bins. Finally, the histogram is normalized to sum up to 1.

Fig. 2. illustrates the generation of HOOF. Binning based on the primary angle, the smallest signed angle between the horizontal axis and the vector helps the histograms to be independent of the (right or left) direction of movement. To make the histogram description scale-invariant, the normalization is done. The significance of each optical flow vector to its corresponding bin is proportional to its magnitude; hence the small noisy optical flow evaluations have little effect on the obtained histogram. Assume a static background, without an optical flow in the background or there is no displacement in the pixels, magnitude-based addition to each bin, can be utilized to calculate the optical flow histogram on the whole image sequence rather than pre-compute a segmentation of the dynamic hand gesture. The number of bins N, is a prime parameter with at least 30 bins per histogram, good recognition results can be achieved. Since HOOF features $h_t = [h_{t;1}, h_{t;2}, \ldots, h_{t;B}]$ at each time instant t are defined at each frame of the video, the actual description is a time series of these histograms $\{h_t\}_{t=0}^{N-1}$. The objective is to perform classification of gestures by comparing these time series. It will not compare the time series directly, it is better to use the temporal evolution of a local feature and integrate with HOOF so that different gestures can be distinguished.

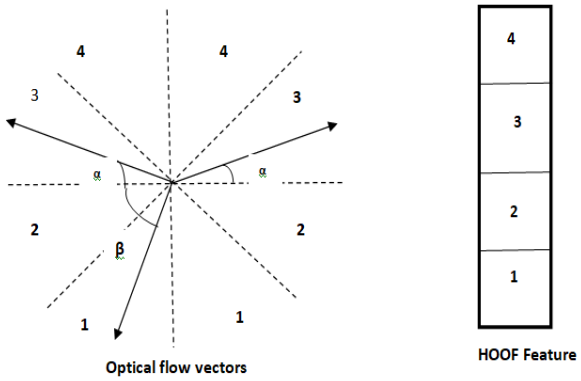

Fig. 2. Formation of histogram with four bins N=4

B. Local Binary Pattern from Three Orthogonal Planes

Motion analysis recent tools are descriptors based on dynamic textures. Based on local binary patterns [10] two spatiotemporal operators were presented for recognizing dynamic textures; Local Binary Pattern histograms from Three Orthogonal Planes (LBP-TOP): XY, XT and YT planes and Volume Local Binary Patterns (VLBP). These features extract the appearance and motion information together. These features are invariant to illumination and translation variations.

The LBP pattern is obtained for XY, XT and YT planes and is represented as LBP − XY, LBP –XT and LBP – YT. These three planes histogram is concatenated into a single histogram. The process of LBP-TOP generation is illustrated in Fig.3. In this, dynamic texture is encoded by the LBP while the motion and shape in two planes of texture are considered, incorporating spatial domain information and two spatial temporal co-occurrence statistics together.

Every pixel in the three planes, a binary code is created by comparing centre pixel with neighboring pixels in a circular neighborhood.

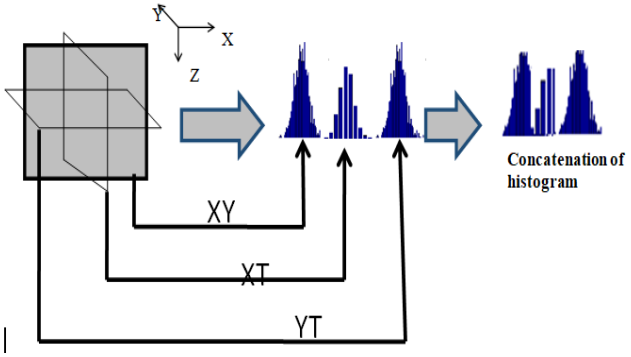

Fig.3. Generation of local binary pattern from three orthogonal planes

Sometimes in LBP-TOP, the radius of three axes X, Y and T, and the number of neighboring points in the XY, XT and YT planes can also be different, that is represented as R_X, R_Y, R and P_{XY},P_{XT},P_{YT}. The resultant LBP-TOP feature is described as $LBP_{P_{XY},P_{XT},P_{YT},R_X,R_Y,R_T}$.

The radius of three axes is same sometimes. Likewise the number of neighboring points in XY, YT and XT planes is also same. Such senarios, we use LBP – $TOP_{P,R}$ for abbreviation where P = P_{XY} = P_{XT} = P_{YT} and R = R_X = R_Y = R_T. Let take an X×Y×T dynamic texture ($x_c \in \{0, \cdots, X-1\}$, $y_c \in \{0,\cdots, Y-1\}$, $t_c \in \{0,\cdots,T-1\}$).

In evaluating $LBP_{P_{XY},P_{XT},P_{YT},R_X,R_Y,R_T}$ distribution for the dynamic texture (DT), the central part is only taken because of sufficiently large neighborhood cannot be utilized on the borders in this 3D space. A histogram of a dynamic texture can be described as

$$H_{i,j} = \sum_{x,y,t} I\{F_j(x,y,t) = i\}, \quad i = 0, \cdots, r_j - 1;\ j = 0, 1, 2. \qquad (2)$$

in which r_j is the number of different labels produced by the LBP operator in the jth plane (j = 0: XY, 1: XT and 2 : YT), $F_{i(x,y,t)}$ denotes the LBP pattern of central pixel (x, y, t) in the j^{th} plane, and

$$I\{A\} = \begin{cases} 1, & \text{if A is true;} \\ 0, & \text{if A is false.} \end{cases} \qquad (3)$$

When the dynamic texture is compared with the different temporal and spatial components, to obtain a coherent description the histograms must be normalized to:

$$N_{i,j} = \frac{H_{i,j}}{\sum_{k=0}^{n_j-1} H_{k,j}} \qquad (4)$$

The LBP-TOP can extract the dynamic texture features effectively. The LBP from XY plane gives the information about appearance and shape. LBP from XT and YT plane gives the information about the horizontal and vertical motion. The three different feature histograms are concatenated together to create a global descriptor, LBP-TOP.

C. Combining HOOF with LBP-TOP

The integration of the two different feature descriptors of the video sequence is,

$$f_{com} = \{f_{motion(1)} * f_{motion(2)}\} = \{w_1.LBP-YT, w_2.LBP-XT, N_{i,j} w_3.LBP-XY\} \qquad (5)$$

where w_i denote to the weights of the particular feature vectors. If the weight is set to be one, then the integration of is simply the concatenation of the different feature vectors.

D. Recognition of hand gestures

Having completed all stages above, the successfully developed feature vector of different hand gestures are tested in the classification stage. Here the classification job is performed via SVM classifier based radial basis function. was created for entire training set. For analysis a ‘leave one group out’ strategy is used.

IV. EXPERIMENTS AND EVALUATION

We have done all the experiments in standard i5 2.7GHz computer with 8 GB RAM. The proposed architecture is tested and analyzed with the selected database such as hand gesture database for human computer interaction [4] and five dynamic hand gestures from RWTHGerman finger spelling database [11].

A. Database

The database consists of dynamic gestures of high quality image sequences created using Senz3D sensor for human computer interaction (HCI) [4]. These images are captured in natural way with uneven background. The videos are created based on the mouse functions such as left click, curser, right click, mouse deactivation and mouse activation. The video sequences were captured in realistic manner with non-uniform background. Two sets of videos are created separately for training and testing purpose. Each set contains five hand gestures performed by various people. The sample frames of different hand gestures from hand gesture database are shown in Fig. 4.

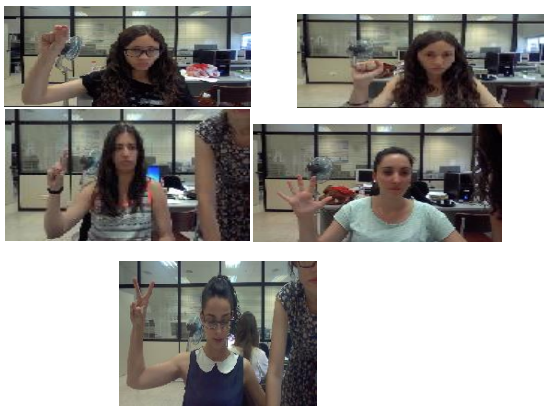

Fig. 4. Sample frames from hand gesture database for HCI

RWTH German fingerspelling database [11] contains 1400 image sequences of 35 gestures showing the letters from A to Z, 'SCH, 1 to 5 numbers and the German umlauts Ä, Ö, Ü. It contain the five dynamic gestures of 700 image sequences. These are captured at 25 frames per second of 320 x 240 pixel resolution with webcam. The videos are taken under different camera viewpoints and cluttered background.
The sample frames of different gesture are shown in Fig. 5.

Fig. 5. Samples of dynamic gestures from RWTH German finger spelling database

B. Evaluation and Comparison

In this section, we first estimate the performance of individual features on hand gesture database. LBP_XT, LBP_XY, LBP_YT features are separately obtained and trained using SVM classifier with radial basis function (RBF) kernel. HOOF features are extracted and trained separately and obtained the result. Then the combined feature descriptor Is trained and tested with the classifier. The complete performance analysis of all the scenarios is shown in Fig. 7.

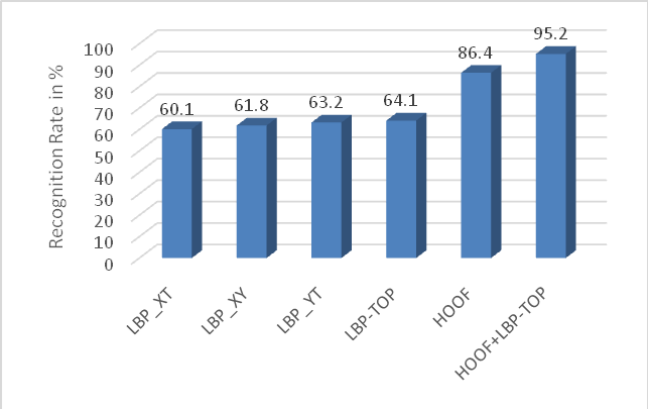

Fig.7. Performance of the individual and combined features on hand gesture database.

From the Fig. 7 we observed that LBP_YT has more contribution on the performance of the system ie., the vertical variations or movements in vertical direction are important in the recognition process. HOOF feature alone got the accuracy of 86.4%. When we combined these two features the performance of the system increased.

Classification accuracy is used as a metric to validate the performance of the system and it can be defined as the number of all correct predictions divided by the total number of gestures in the dataset. Fig. 8 shows the accuracy of different hand gestures in hand gesture database tested on proposed framework. The accuracy obtained for the architecture is 0.952.

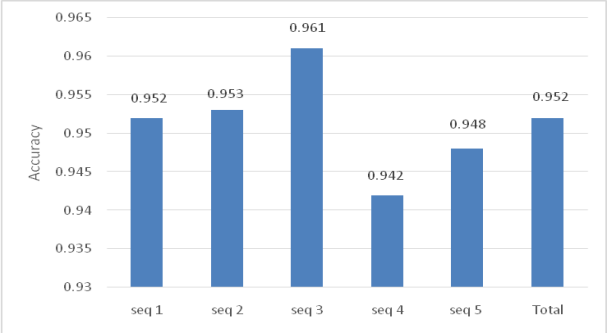

Fig.8. Performance of various hand gestures in hand gesture database for HCI based on accuracy

Next the performance of the framework is compared with the other state of the art methods and is shown in Table 1. The temporal information combined with local and global data produces volumetric spatiograms of local binary pattern (VS-LBP) obtained an accuracy of 0.927 [4]. AI-Berry et al., 2015 [17] introduced a combination of features that fuses moments and directional wavelet Local Binary Pattern (LBP) and result in an accuracy of 0.914. In our earlier work [15], we developed a hand gesture recognition based on an optical flow and pyramid histogram of gradients (PHOG) and obtained an accuracy of 0.946. We developed a SIFT flow based framework and achieved an accuracy of 0.976 [16]. But it has the limitation of using optimization technique. Our proposed framework gives better results when compared with other existing methods.

Table 1: Performance comparison of the proposed framework with other methods using Hand gesture database

Method	seq_1	seq_2	seq_3	seq_4	seq_5	seq_6	Mean Accuracy
VS-LBP [Maqueda et al.,2015]	0.951	0.962	0.939	0.931	0.801	0.959	0.927
LBP+ Moments [AI-Berry et al., 2015]	0.942	0.918	0.937	0.923	0.851	0.923	0.915
PHOG_TOP+ Optical flow [Suni et al., 2020]	0.967	0.971	0.942	0.939	0.892	0.962	0.946
Dense SIFT flow based framework [Suni et al., 2020]	0.981	0.978	0.969	0.976	0.982	0.971	0.976
Proposed framework	**0.967**	**0.971**	**0.945**	**0.953**	**0.912**	**0.962**	**0.952**

The same set of experiments are repeated on RWTH gesture database by Drenw [11]. The confusion matrix produced for the five dynamic hand gestures ('J', 'Z', Ä, Ö and Ü) on this database is shown in Table 2. The recognition rate obtained is 90% for this database using the developed framework.

Table 2. Confusion matrix obtained for five dynamic gestures in RMTH gesture Database

GESTURE	J	Z	Ä	Ö	Ü	RECOGNITION RATE(%)
J	17	0	1	2	0	85
Z	0	17	1	0	2	85
Ä	0	1	18	0	1	90
Ö	0	1	1	18	0	90
Ü	0	0	1	1	18	90

Table 2. Performance comparison of proposed framework With other state of the art methods

Method	Description	Error rate
Dreuw et al., 2006	Thresholding + camshift tracking	35.7%
Abdalla, 2013	motion and contour shape features of hand	27.6%
Suni et al., 2020	combination of Pyramid histogram of gradients and optical flow features	11%
Proposed architecture	**HOOF+ LBP_TOP**	**10%**

Table 2 demonstrates the performance comparison of the introduced method with existing approaches on RMTH German fingerspelling database in terms of error rate. Dreuw et al., 2006 [11] created a hidden Markov model based on appearance features to recognize the hand gestures and was obtained an error rate of 35.7%. Abdalla, 2013 [18]achieved an error rate 27.6% by using motion and contour shape features of the hand. In our earlier work [15] using a multiple feature based framework obtained an error rate of 11%. The proposed framework achieved an error rate of 10%, which better than other state of art methods.

V. CONCLUSION

A framework for dynamic hand gesture recognition has been proposed by fusing the histogram of oriented optical flow (HOOF) and local binary pattern from three orthogonal planes (LBP-TOP). The framework is proved to be good for dynamic hand gesture recognition under different camera viewpoint and cluttered background. Experimental analysis showed that vertical movements of hand gestures play a vital role in the performance of the system. The framework gives the recognition rate of 95.2%. This approach is robust to photometric variations, scale and direction of motion.

ACKNOWLEDGMENT

The authors would like to thank Kerala University Fellowship (order No:AcE1.A4/714/ECE/33243-1/2014) for providing funds to the research.

REFERENCES

1. Siddharth S. Rautaray and Anupam Agrawal, " Vision based hand gesture recognition for human computer interaction: a survey", Artificial intelligence Review", vol. 43, pp 1- 54,2015.
2. Joyeeta Singha, Amarjit Roy and Rabul Hussain Laskar, "Dynamic hand gesture recognition using vision-based approach for human–computer interaction" , Neural Comput & Applic, vol. 29, pp. 1129 -1141, 2018.
3. Dan Xu, Xinyu Wu, Yen-Lun Chen and Yangsheng Xu, " Online Dynamic Gesture Recognition for Human Robot Interaction", J Intell Robot Syst, vol. 77, pp 583-596, 2015.
4. Ana I. Maqueda, Carlos R. del-Blanco, Femando Jaurequizar and Narciso Garcia, " Human-computer interaction based on visual hand-gesture recognition using volumetric spatiograms of local binary patterns", computer Vision and Image Understanding, vol. 141, pp 126-137, 2015.
5. Xijian Fan and Tardi Tjahjadi, " A spatial – temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences", Pattern Recognition, vol. 48, pp. 3405-3416, 2015.
6. Jinqing Zheng, Zhiyong Feng, Chao Xu, Jing Hu and Weimin Ge, " Fusing shape and spatio-temporal features for depth based dynamic hand gesturerecognition" Multimedia Tools andApplications, vol. , pp. 1-20, 2016.
7. A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In IEEE International Conference on Computer Vision, pages 726–733, 2003.
8. D. Tran and A. Sorokin. Human activity recognition with metric learning. In European Conference on Computer Vision, 2008.
9. J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of humanactioncategoriesusingspatial-temporalwords. International Journal of Computer Vision, 79:299–318, 2008.
10. G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions", IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 6, pp. 915 – 928, Jun, 2007.
11. P. Dreuw, T. Deselaers, D. Keysers and H. Ney, "Modeling image variability in appearance-based gesture recognition," In ECCV Workshop on Statistical Methods in Multi-Image andVideo Processing, pp. 7-18, 2006.
12. S. S. Suni and K Gopakumar," A real time decision support system using head nod and shake", IEEE Int. Conf. on Circuit, Power and Computing Technologies, 2016
13. Jinqing Zheng, Zhiyong Feng, Chao Xu, Jing Hu and Weimin Ge, " Fusing shape and spatio-temporal features for depth based dynamic hand gesturerecognition" Multimedia Tools and Applications, vol. , pp. 1-20, 2016.
14. G. Willems, T. Tuytelaars, and L. J. V. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In European Conference on Computer Vision, 2008.
15. S S Suni and K Gopakumar, "Fusing pyramid histogram of gradients and optical flow for hand gesture recognition", International Journal of Computational Vision and Robotics, Inderscience, 10(5), pp. 449-464, 2020.
16. Suni S S and K Gopakumar, "Dense SIFT–Flow based Architecture for Recognizing Hand Gestures", Advances in Science, Technology and Engineering Systems Journal, ASTES, 5(5), pp. 944-954, 2020.
17. Maryam Nabil AI-Berry, Mohammed A. M. Salem, Hala Mousher Ebeid, Ashraf S. Hussein and Muhammed F. Tolba, 'Fusing directional wavelet local binary pattern and moments for human action recognition', IET ComputerVision, vol.10, Issue. 2, pp. 1-10, 2015.
18. Abdalla, M.S. and Hemayed, E.E, "Dynamic Hand Gesture Recognition of Arabic Sign Language using Hand Motion Trajectory Features", Global Journals Inc. (USA), vol. 13, 2013.

AUTHORS PROFILE

Suni S S, is a PhD student in Electronics & Communication at the University of Kerala, India. She received her Master degrees in Optoelectronics & Optical Communication from the Optoelectronics Department at the same university in 2007. Her current research areas are image and video processing, Human – Computer Interaction and Computer vision.

Gopakumar K, received his Ph.D. (Faculty of Engineering and Technology) in Chaos Theory and Applications from University of Kerala (2012). He is currently working as Professor, Department of ECE, TKM College of Engineering, Kollam, Kerala. He is the member of Indian Society for Technical Education (ISTE), Fellow of Institute of Electronics & Telecommunication Engineering (IETE) and The Institution of Engineers India (IEI). He published many papers in International Journals, International Proceedings and National Proceedings and also authored several text book in engineering.