ACTION RECOGNITION IN LOW RESOLUTION VIDEOS USING FO-SVM

K.RangaNarayana[1]

Research Scholar[1], Department of IT [1], GITAM Deemed to be University[1], Visakhapatnam, India.
katakam916@gmail.com[1]

G.Venkateswara Rao[2]

Professor[2], Department of CSE[2], GITAM Deemed to be University[2], Visakhapatnam, India.
dr.vrgurrala@gmail.com[2]

Abstract

The majority of extant in recognition of action research is focused on high-quality videos with clearly apparent activity. The actions in videos are collected at a variety of resolutions in real-world surveillance situations. Most actions take place at a distance with a low resolution, and identifying such events is a difficult task. In this paper, we take a look at the effect of low video quality on human activity location from two points: recordings that are weakly gathered topographically (low level resolution) and transiently (lower frame rate), and packed recordings with movement obscuring and artifacts. To recognize the action four main steps are considered. First one is background subtraction using LPB operator. Second, the features are extracted using histogram of gradients (HOG) and Histogram of optical flow (HOF) algorithm which is used to estimate the motion of a person and eigen value algorithm which is used to recognize the person. These features are fused together and perform Firefly optimization (FO) technique to obtain optimized features. Thirdly, generate a code book for feature encoding. In feature encoding, the process is performed using bag of words. Finally action recognition is done using Support vector machine (SVM) classifier. The experimental results are performed on low resolution datasets like VIRAT dataset, KTH dataset and Soccer dataset. The result obtained using VIRAT is 91.46%, KTH is 92.40 % and Soccers is 90.51% in terms of accuracy.

***Keywords*: Low resolution videos; Eigen Vectors; Firefly Optimization; SVM**

1. Introduction

Human activity acknowledgment in video is a functioning space of examination, with some certifiable applications going from automated video observation, video mining and recovery, and collaborative computer games. In unconstrained recordings, human activities are caught with run of the typical picture varieties like appearance, scale and view present, to really testing issue, for example, enlightenment change, impediment, shadow, and camera movement. One moderately neglected issue is relating the nature of recordings. As most examination in real life acknowledgment depend with the understanding that video information is of superior quality i.e. HD with insignificant sign commotion, many suggested approaches have found to functioned admirably under such immaculate conditions. Notwithstanding, the vast majority of these recordings are not practical for constant video preparing, information web based and portable applications, because of the computational overhead that most techniques bring about. Famous methodologies for nonexclusive picture grouping have been reached out for use in video arrangements, to a decent proportion of achievement. Specifically, sack of-words (or pack of-highlights) based strategies have exhibited promising outcomes for the undertaking of activity acknowledgment [1]–[3]. Indeed, even with these triumphs, the portrayal of neighborhood areas in recordings is as yet an open issue in research. Accordingly, an assortment of spatio-worldly highlights has likewise been considered in writing to more readily address video information. Numerous mainstream works [4-5] favor using inclination and stream data to portray the shape and movement that lies in the video. The utilization of surfaces, in any case, is more uncommon [6-7], however there are numerous advantages that can be utilized.

In [8] the original version for continuous surveillance is a new dataset i.e. VIRAT and is a large scale, supplied nine down sampled variants of the data, which consist of geographical scales of three types and temporal frame rates of three types is used. The authors emphasized that this is a "relatively uncharted region" and that "understanding how existing methods would act differently" is critical. Thus, this inspires this work to examine the ability of perceiving activities under these difficult conditions. Motivated by the known benefits of different highlights and the undeniable absence of activity acknowledgment work in bad quality recordings, we plan to research and present an achievable way to deal with this issue. In this paper, the use of shape and movement highlights to expand the heartiness of perceiving human activities under various conditions is suggested. The work is assessed for inferior quality recordings. We test the suggested technique with a series of comprehensive experiments on three well-known benchmark action datasets of varying nature: the VIRAT Dataset, the KTH Dataset, and the Soccer Dataset, which comprises basic actions in a controlled setting.

2. Dataset Used

2.1. *VIRAT Dataset*

TinyVIRAT is a low-resolution action recognition system. Tiny VIRAT's videos are realistic and were taken from surveillance videos in the VIRAT dataset [9]. This dataset have many action based video clips with multiple operations, which adds to the difficulty. The collection contains about 13K video samples from 26 distinct activities, all of which are recorded at 30 frames per second. The duration of the activity varies depending on the sample, with an average duration of about 3 seconds. It comprises a variety of low-resolution films ranging in size from 12x12 pixels to 132x132 pixels, taking an average size of 80x80 pixels. The videos in the suggested collection are inherently lower resolution and depict real-world difficulties.

2.2. *KTH Dataset*

KTH [10] is one of the most often used dataset for performing human action recognition experiments. Different actions video clips such as running, person jogging, hand waving by a person, clapping of hands, walking movements of a person and boxing. These actions are performed by 25 people in various environmental conditions like indoors, outdoors, indoor with different dressing and outdoor with different dressing. The total number of video clips available in this dataset is around 599. Every clip present with a time bound of 10-15 sec and with a frames rate of 2 frames per second and the resolution termed to be 160 120 pixels size.

2.3. *Soccer Dataset*

The name implies sports; in this collection, the focus is on sports footage acquired for academic purposes. A database based on footballs is compiled and made available to researchers in the form of short, event, and tracking (SSET). The dataset comprises of 350 soccer movies from various soccer games played across the world over the course of 282 hours. This database is divided into three sections. One is brief and consists of five types and two types. The second category is event/story, which includes 11 events and 15 narrative types. The third component is the scope of players, which consists of links, the breadth and length of the scope [11].

3. Firefly Optimization

Xin-She Yang created the Firefly Algorithm (FA) in late 2007 and early 2008 at Cambridge University [12-13], which was based on the flashing patterns and behaviour of fireflies. In recent years, the evolution of the Firefly Algorithm has become increasingly significant, and numerous research projects have been carried out using the Firefly Algorithm. The findings produced through these optimization approaches were deemed accurate, and numerous articles were published as a consequence.

Fireflies are unisex, which means that any individual firefly will be drawn into various fireflies regardless of their opposite sex. Attractive nature and brightness are closely related to each other, increase in separation of flies causes decrease in attractive nature and brightness. Similarly, for any two burning fireflies, the less brilliant one will attract the more brilliant one. If there is no more brilliant one than a certain firefly, it will travel at random. The scene of the target work controls the brightness of a firefly [13]. The intensity of light and the attractiveness are proportional to each other, the attractiveness β with r distance is given as,

$$\beta = \beta_0 e^{-\gamma r^2} \quad (1)$$

Where β_0 is attractiveness at r=0. The firefly '*i'* is attracted to firefly 'j' which is more brighter and move towards 'j' and the same is determined by

$$x_i^{t+1} = x_i^t + \beta_o e^{-\gamma r_{ij}^2}\left(x_j^t - x_i^t\right) + \alpha_t \epsilon_i^t \quad (2)$$

Where the second term $\beta_o e^{-\gamma r_{ij}^2}\left(x_j^t - x_i^t\right)$ is concern based on the attraction. The third term i.e. $\alpha_t \epsilon_i^t$ is termed as randomization with α_t being the parameter of randomization, and ϵ_i^t is a vector of random numbers drawn from a distribution sources which may be uniform or gaussian at time *t*. If β_0= 0, it becomes a regular pattern for determining the optimized value. If γ = 0, FA in generally termed to be swarm based optimization [12].

Algorithm:

The firefly algorithm is implemented and process is stated below:

Step1. Initialization

Step2. The population of fireflies are gives as {x_1, x_2, …….. , x_n}

Step3. Calculate brightness value using cost function for assigned firefly

Step4. Firefly intensity is given as {I_1,I_2, ……. , I_n}

Step5. Update the step of each firefly

Step6. Ranking of fireflies and finding current best

Step7. Moving firefly i towards other firefly which is brighter

Step8. Update the solution set

Step9. Stop when result obtained; otherwise go to Step 2.

4. Methods

For detecting moving people optical flow methods has more importance, whereas for identifying static pictures eigen vector is a one of the significant approach. The optimization approach aids in getting the optimal features, and SVM classification is utilized to detect human activity in low-resolution movies.

4.1. *Input Data*

The low resolution videos are obtained from various datasets discussed in section 2. The video which is considered for processing are converted into frames.

4.2. *Background* Subtraction

The following stage is to deduct the foundation from the gray scaled video edges and concentrates just moving object data. This deduction step should be possible by taking the initial not many casing of the foundation and subtract it when the moving object was distinguished.

A local binary patter technique is adopted for performing the background subtraction process [14]. Some of important properties that LBP operator has are termed as invariance of gray scale, non-parametric, Invariant illumination, simple for computing, and more discriminative. The first version has eight neighbour pixels but this may be readily modified to accommodate a large neighbourhood which is circular with 'n' number of pixels. The sign of the difference between the centre pixel and its P neighbours is employed as a threshold in LBP to calculate the binary number with P-bit, resulting in two discrete decimal values for the patterns in binary.

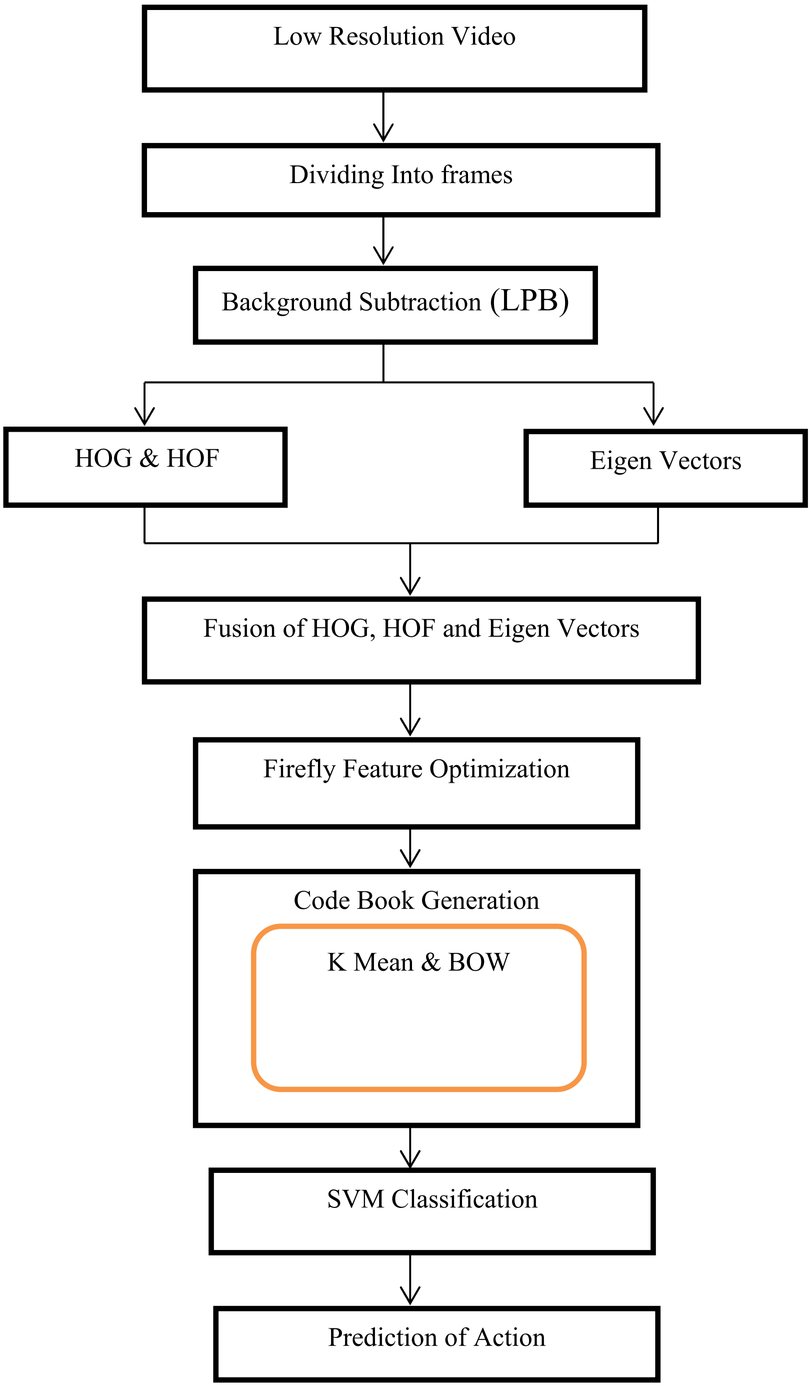


Fig 1. Framework of Proposed Method

4.3. *Feature Extraction*

After the process of background subtraction, these frames are used for extracting features. The features are extracted using optical flow algorithm for estimation of motion and eigen values features are extracted for recognizing the person.

4.3.1. *HOG and HOF features*

In [15] presented the HOG descriptor as a human detection method. To extract features, they employ a dense grid of histograms of oriented gradients computed over a block. The optimization approach is provided the characteristics. We utilize the HOG method, but with a different feature space: we use 8 × 8 pixel cells without grouping them into blocks, and we use a 40 × 40 pixel window. Each object is scaled to that size. The original HOG was only used on one picture at a time.

Among the most effective methods for recognizing moving individuals in videos is to observe the differences between two or more consecutive and adjacent frames in the video.. A pattern of optical flow is thought to be continuous motion of objects. This pattern is caused by the observers' and objects' relative movements [16]. Optical flow fields are then utilized to generate an orientation histogram. The orientation histogram was represented using an 8-bin format. Flow is heavily influenced by the frame rate of the original video. As a result, a low frame rate produces a lower quality HOF feature. Eigen characteristics are extracted from each frame and preserved for later identification of the individual in the video. HOG, HOF, and Eigen characteristics are retrieved and fused together to provide a higher rate of accuracy in detecting a person's action in low quality footage.

4.4. *Optimization of Features*

The combined features are given to optimization technique. In this we used firefly optimization technique for obtained the best features. The process of Firefly optimization is discussed in section 3. Based on the process of optimization the accurate features that are required to identify the human action is evaluated.

4.5. *Code Book Generation*

The code book will be made dependent on the advanced highlights. A video grouping is characterized as an assortment of extracted features. The attributes are quantized into a bunch of visual words known as a codebook, and the video is communicated by a recurrence histogram of these word vectors. Pack Of-Words (BOW) is one of the portrayal strategies, every one of which makes its own codebook to encode the element descriptors. Codebooks can be worked from an interesting location metric or a combination of a few element descriptors joined at the descriptor level.

BOW (Bag-of-Words): Standard k-means clustering is used to produce codebooks. For stability, we provisionally set the number of word vectors at V = 4000, which has been displayed in a few examination [17] to deliver good results. Naturally, vector quantization (VQ) is used for feature encoding, this is a form of hard general administrative on the nearest Euclidean distance. If a feature is closer to the cluster c centroid than any other centroids, it is allocated to cluster c. To improve the precision in clustering, we repeatedly run k-means for 8 times by considering previous rounds and keep the result with the lower error.

4.6. *SVM Classification*

The support vector machine is one of the most often used classifiers. The SVM will be trained using the optimal dataset features acquired by utilizing FO to build an SVM model. Around 70% of the data is utilized for training, with the remaining 30% used for testing. For classification, a linear SVM is employed. We set the class weight w for the SVM hyper-boundaries to be conversely corresponding to the measure of tests in each class, with the end goal that both positive and negative classes contribute similarly to the loss function. At last, the strategy will be resolved.

5. Experimental Evaluation

The suggested technique is reviewed using the KTH Dataset, Soccer Dataset, and VIRAT Dataset. Three datasets are commonly used in person recognition in low-resolution movies. The dataset is structured in order to produce experimental findings. The dataset is made up of several films of varying durations. These movies were captured with lower-resolution cameras. The proposed approach is utilized to evaluate the parameters. The human action recognition of several datasets is depicted here.

5.1. *KTH Dataset*

Here we consider different types of low resolution videos from KTH dataset for performing experimental results. Different action videos are given as input for identifying the action of human and provide the results

Case1.

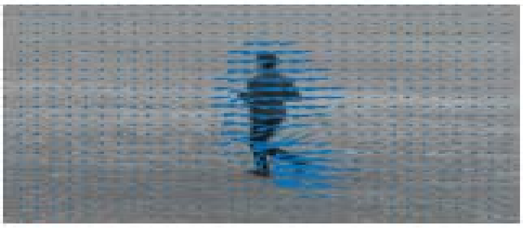


Fig 2. Frame extracted for given input video

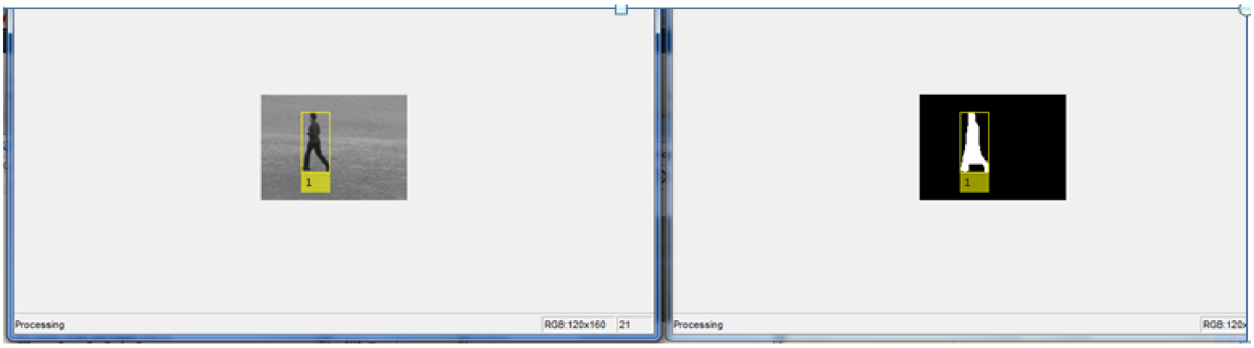
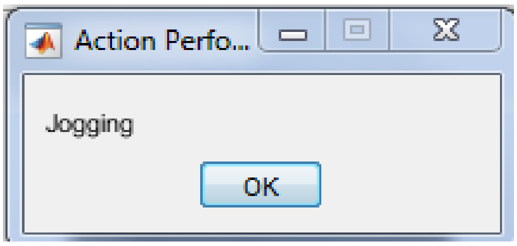


Fig 3. Processing of frames



Jogging

Fig 4. Prediction of Action

Case2.

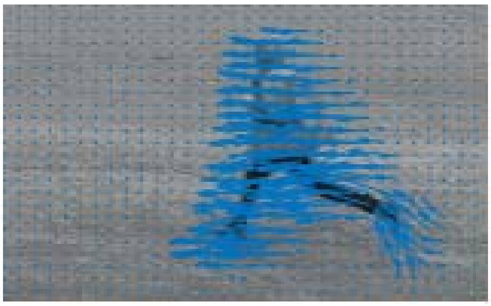


Fig 5. Frame Extracted from Input video

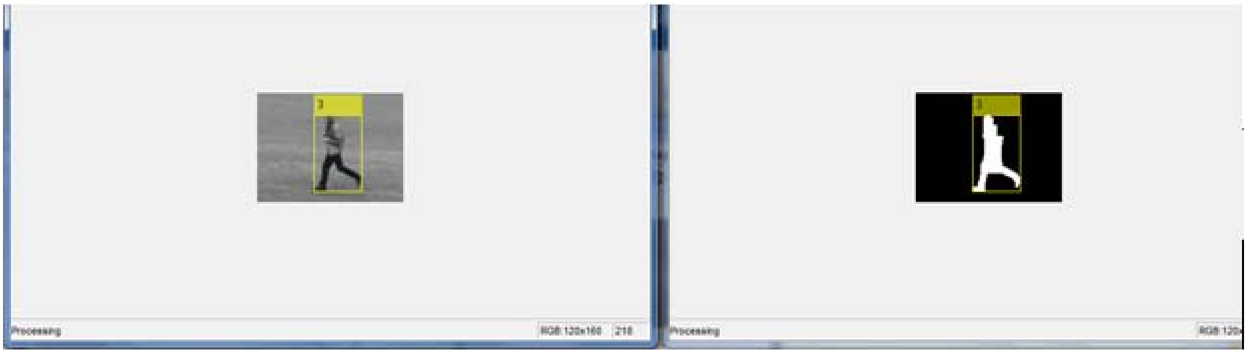
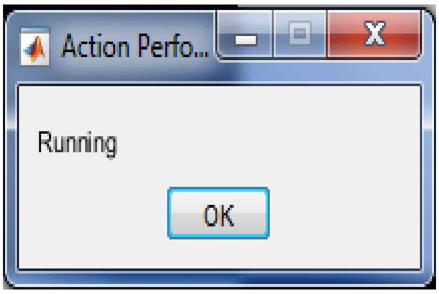


Fig 6. Processing of frames



Running

Fig 7. Prediction of action

Case3.

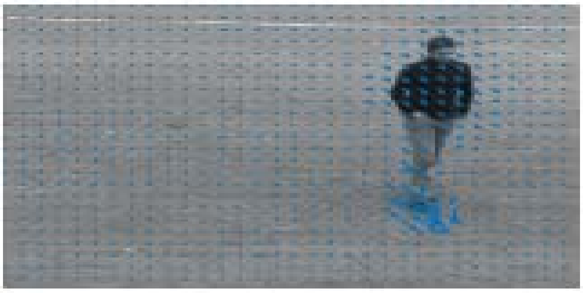


Fig 8. Frame Extracted from input video

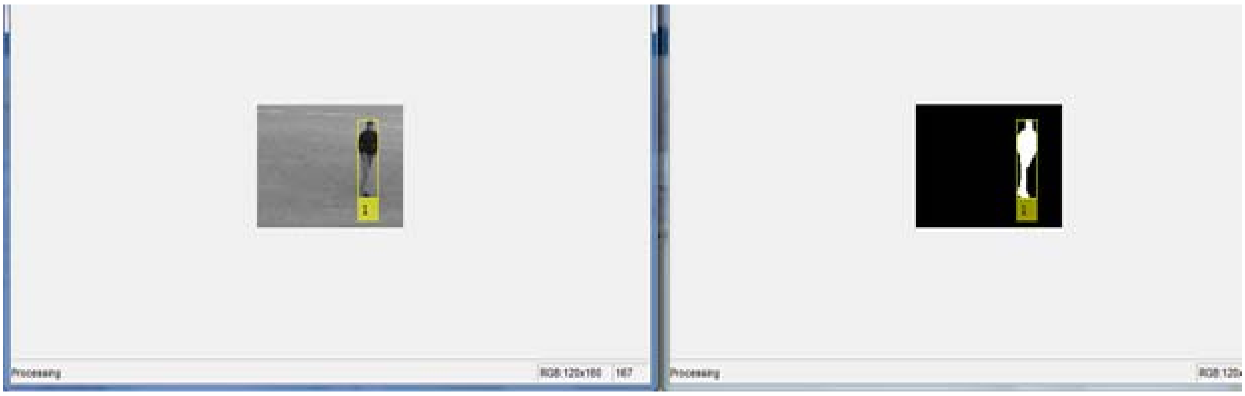
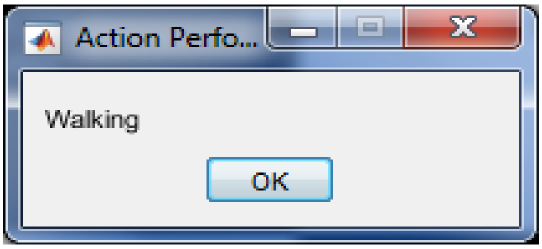


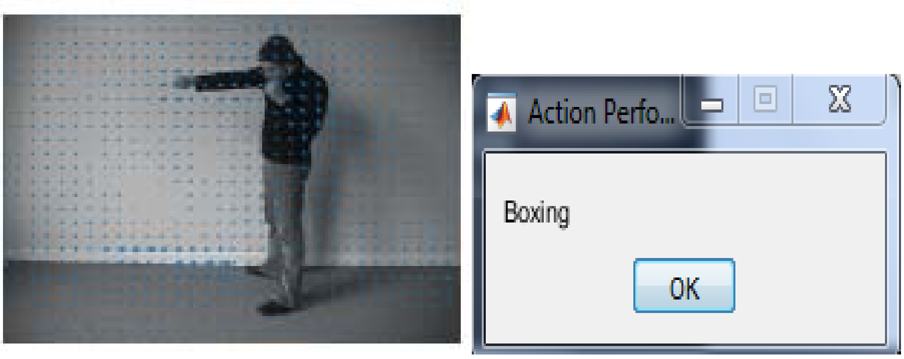
Fig9. Processing of frames



Walking

Fig 10. Prediction of output

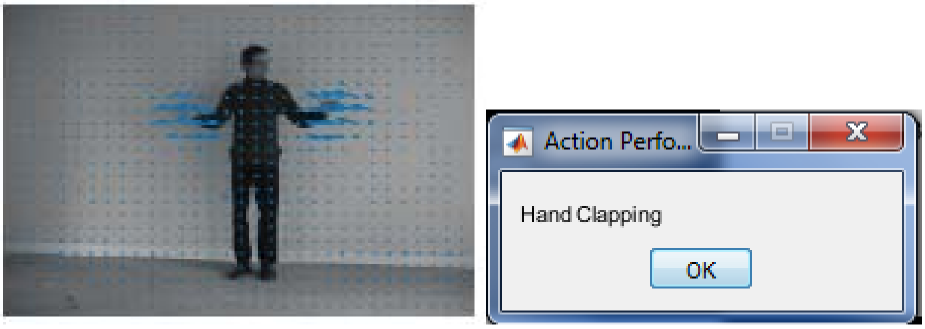
Case4.



Boxing

Fig 11. Frame extracted and prediction of action is boxing

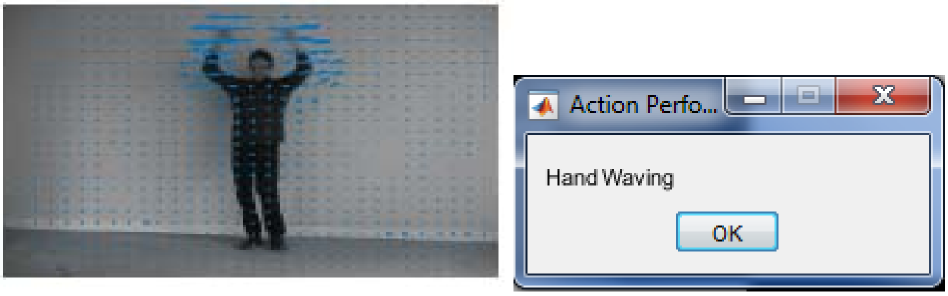
Case5.



Hand Clapping

Fig 12. Frame extracted and prediction of action is hand clapping

Case6.



Hand Waving

Fig 13. Frame extracted and prediction of action is hand clapping

5.2. *Soccer Dataset*

Here we consider different types of low resolution videos from soccer dataset for performing experimental results. Different action videos are given as input for identifying the action of human and provide the results.

Case1.

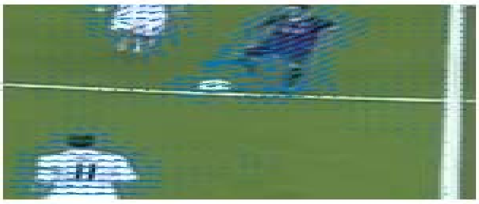


Fig 14. Frame extracted from input video

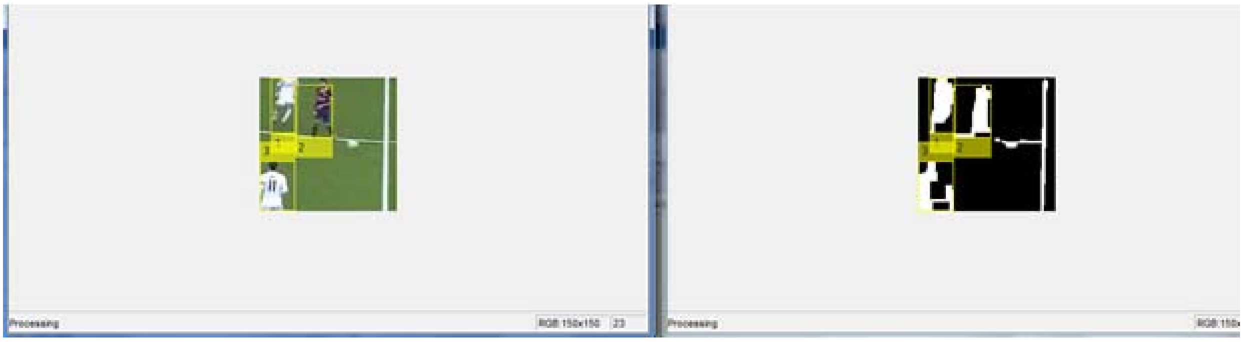
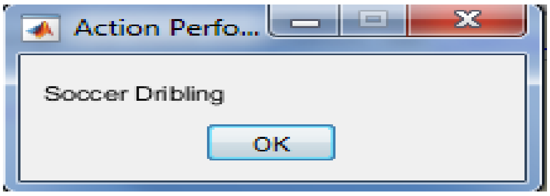


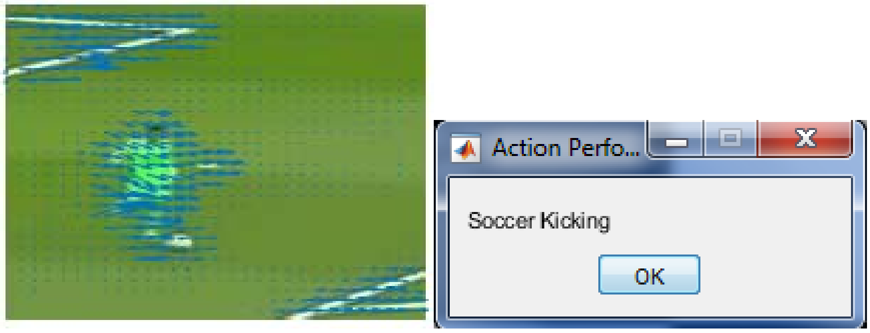
Fig 15. Processing of frames



Soccer Dribling

Fig 16. Prediction of action

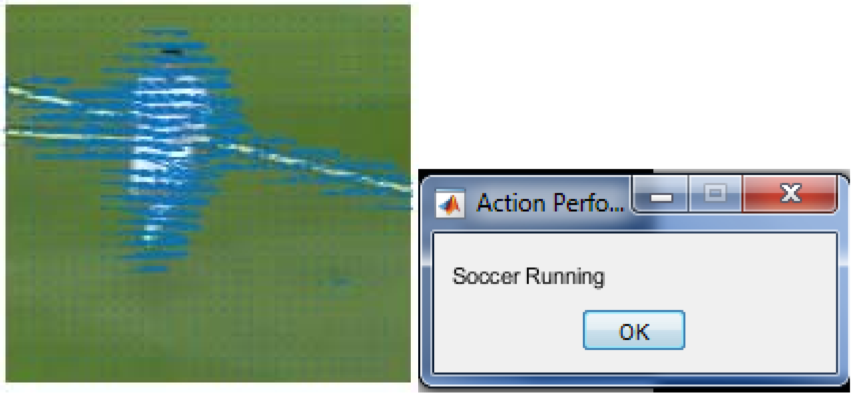
Case2.



Soccer Kicking

Fig 17. Frame extracted and prediction of action is soccer kicking

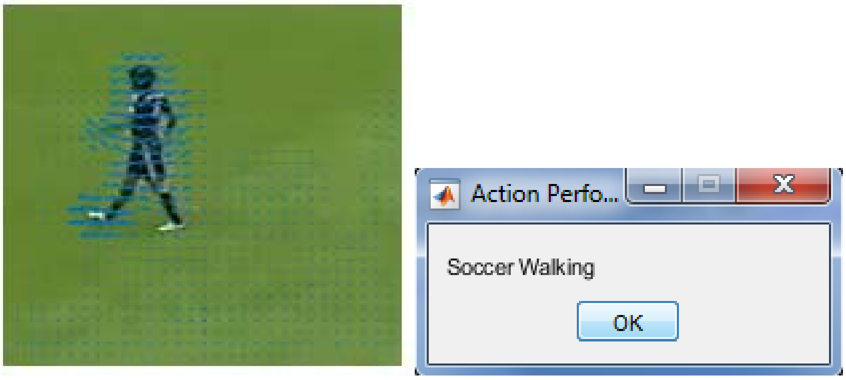
Case3.



Soccer Running

Fig 18. Frame extracted and prediction of action is soccer running

Case4.



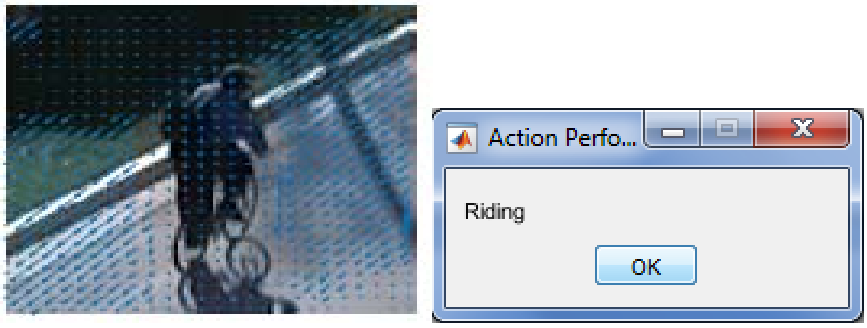
Soccer Walking

Fig 19. Frame extracted and prediction of action is soccer walking

5.3. *VIRAT Dataset*

Here we consider different types of low resolution videos from VIRAT dataset for performing experimental results. Different action videos are given as input for identifying the action of human and provide the results

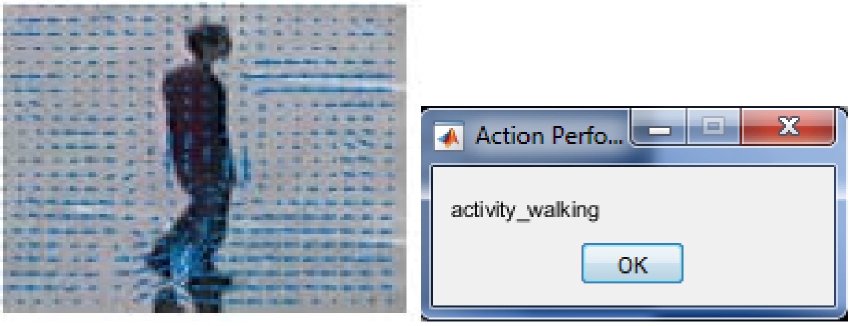
Case1.



Riding

Fig 20. Frame extracted and prediction of action is riding

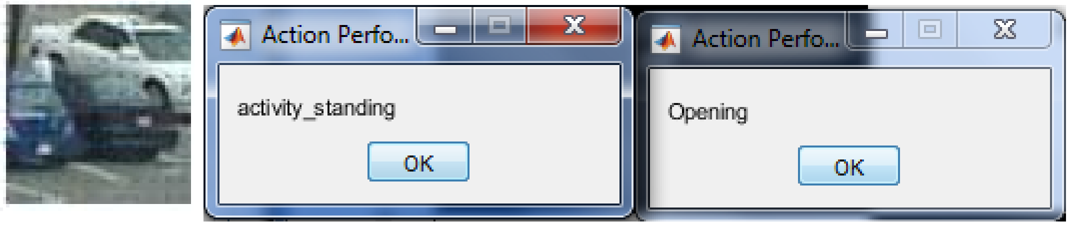
Case2.



activity_walking

Fig 21. Frame extracted and prediction of action is walking

Case3.

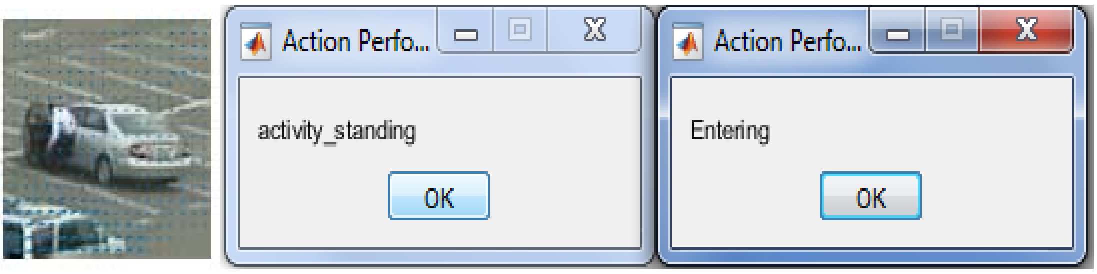


activity_standing

Opening

Fig 22. Frame extracted and prediction of action is standing and opening door

Case4.



activity_standing

Entering

Fig 23. Frame extracted and prediction of action is standing and entering into car

5.4. *Evaluation Metrics*

The assessment metrics are used to distinguish between the conventional and suggested techniques. Table 1 shows the terms that will be used to assess the proposed effort. The dataset undergoes training and testing. First, SVM is trained, and then a model is built, one for each dataset. For the remainder of our trials, we put our person recognition algorithms to the test on a fair fraction of videos from the VIRAT, Soccer, and KTH test sets.

The results obtained using Firefly optimization and without using optimization techniques are compared with other optimization technique and the results are shown in table1 and table2

		KTH Dataset	Soccer Dataset	VIRAT Dataset
Accuracy	SVM[19]	85.29	84.15	82.63
	Proposed method	**86.05**	**87.19**	**84.34**
Detection Rate	SVM[19]	83.78	82.61	81.67
	Proposed method	**84.57**	**85.32**	**83.07**
False Detection rate	SVM[19]	14.37	15.56	18.16
	Proposed method	**13.57**	**11.77**	**16.92**
FPPI	SVM[19]	13.28	14.39	16.48
	Proposed method	**13.08**	**14.29**	**16.32**
Miss Rate	SVM[19]	16.21	17.38	18.32
	Proposed method	**12.54**	**11.00**	**14.46**
F1 Score	SVM[19]	84.69	83.514	81.754
	Proposed method	**85.48**	**86.75**	**83.64**
Precision	SVM[19]	85.628	84.431	81.836
	Proposed method	**86.42**	**88.22**	**84.23**

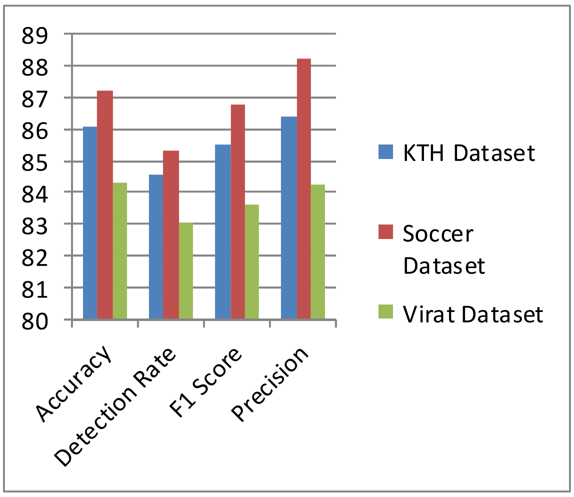
Table 1. Evaluation of Datasets without using optimization technique

The use of optmization technique helps in improving the rate of accuracy and other parametric values for person recogniton in low resolution videos. The parameters values obtained using code book generation and FO-SVM classification are shown in table 2. The values obtained using table2 do include the process of optimization.

		KTH Dataset	Soccer Dataset	VIRAT Dataset
Accuracy	PSO-OFA[19]	90.512	89.56	88.61
	Proposed method	**91.46**	**92.40**	**90.51**
Detection Rate	PSO-OFA[19]	90.02	88.56	88.48
	Proposed method	**90.50**	**91.84**	**89.70**
False Detection rate	PSO-OFA[19]	9.98	10.37	12.57
	Proposed method	**7.58**	**7.78**	**9.58**
FPPI	PSO-OFA[19]	9.04	9.50	11.27
	Proposed method	**9.01**	**8.15**	**10.29**
Miss Rate	PSO-OFA[19]	9.98	11.43	11.51
	Proposed method	**7.05**	**7.07**	**8.74**
F1 Score	PSO-OFA[19]	90.12	89.08	89.95
	Proposed method	**91.14**	**92.03**	**90.05**
Precision	PSO-OFA[19]	90.23	89.62	89.425
	Proposed method	**92.41**	**92.21**	**90.41**

Table 2. Evaluation of Datasets using optimization technique

The changes observed in proposed method for various datasets are shown using bar graphs. Here, the accuracy, decision rate, F1 score, Precison comparison, False detection rate, FPPI and miss rate comparison is shown in fig 23 and fig 24.



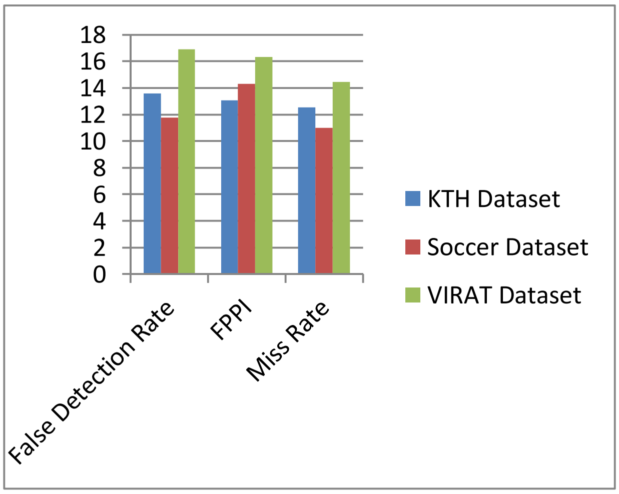
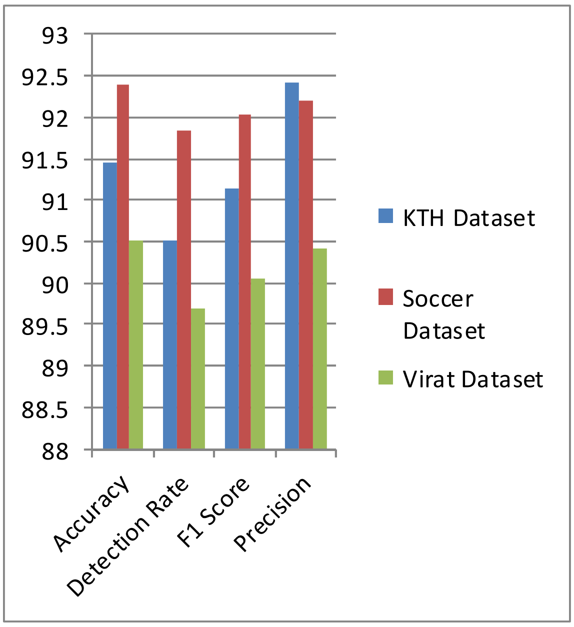


Fig 24. Comparison results obtained using HOG&HOF and EigenVectors with codebook generation.

The comparison shown in figure 23 are evaluated using HOG&HOF and Egienvectors features and classified using support vector machine. Among the three datasets KTH dataset results having an accuacy of 85.10. The other parameters like F1 score and precison are better for KTH dataset.



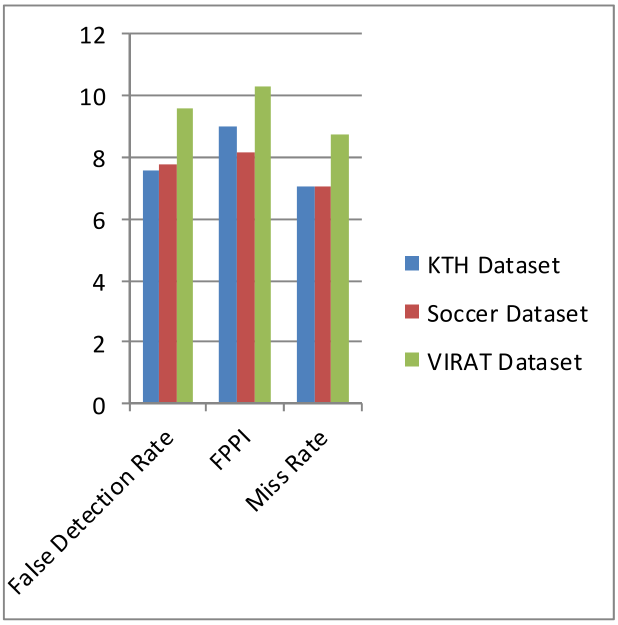


Fig 25. Comparison results obtained using proposed FO-SVM.

From fig 24 it is shown that proposed FO-SVM techniques performs well in terms of all the parameters evaluated compared to SVM. The accuracy obtained for KTH dataset using optimization technique is 92.40% which is 5% higher than without optimization technique.

6. Conclusion

In this research, we look at how low video quality affects human action recognition. The suggested technique has been developed and shown on a variety of low resolution video datasets such as VIRAT, KTH, and Soccer. The HOG and HOF features, as well as the Eigen vector features, are retrieved and combined. The optimization approaches aid in the optimization of the characteristics. The procedure of generating the code book is completed. The SVM classifier produces decent results. Our suggested approach identifies actions such as walking, running, hand waves, jogging, and etc. We perform experiments on three different benchmark datasets and demonstrate that the proposed approach leads to a better action recognition performance in terms of accuracy and other parameters which are evaluated. The accuracy obtained using proposed method is 90.51% for VIRAT dataset, 91.46% for KTH dataset and 92.40% for Soccer dataset.

References

[1] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in *IEEE CVPR*, 2008, pp. 1–8.

[2] J. C. Niebles, H. Wang, and L. Fei-Fei, “Unsupervised learning of human action categories using spatial-temporal words,” *IJCV*, vol. 79, no. 3, pp. 299–318, 2008.

[3] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, “Evaluation of local spatio-temporal features for action recognition,” in *BMVC*, 2009, pp. 124–1.

[4] P. Scovanner, S. Ali, and M. Shah, “A 3-dimensional sift descriptor and its application to action recognition,” in *Proceedings of the 15thinternational conference on Multimedia*. ACM, 2007, pp. 357–360.

[5] H. Wang, A. Klaser, C. Schmid, and C.-L.Liu, “Action recognition by dense trajectories,” in *Computer Vision and Pattern Recognition (CVPR),2011 IEEE Conference on*. IEEE, 2011, pp. 3169–3176.

[6] V. Kellokumpu, G. Zhao, and M. Pietik¨ainen, “Human activity recognition using a dynamic texture based method.” in *BMVC*, 2008.

[7] R. Mattivi and L. Shao, “Human action recognition using lbp-top as sparse spatio-temporal feature descriptor,” in *Computer Analysis ofImages and Patterns*. Springer, 2009, pp. 740–747.

[8] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen *et al.*, “A largescalebenchmark dataset for event recognition in surveillance video,” in *IEEE CVPR*. IEEE, 2011, pp. 3153–3160.

[9] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee,S. Mukherjee, J. Aggarwal, H. Lee, L. Davis et al., “A large-scale benchmark dataset for event recognition in surveillance video,” in CVPR 2011. IEEE, 2011, pp. 3153–3160.

[10] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: a local svm approach,” in *Int. Conf. on Pattern Recognition*, vol. 3, 2004, pp. 32–36.

[11] K.Ranganarayana And G. VenkateswaraRao, 2020, “A Study on Approaches For Identifying Humans In Low Resolution Videos”, International Journal of Advanced Research in Engineering and Technology (IJARET).Volume:11,Issue: 12,Pages:1665-1679.

[12] Yang, X.S.: Nature-Inspired Metaheuristic Algorithms. Luniver Press, UK (2008)

[13] Yang, X.S.: Firefly algorithms for multimodal optimization. In: Stochastic Algorithms: Foundations and Applications, SAGA 2009. Lecture Notes in Computer Sciences, vol. 5792, pp. 169–178 (2009)

[14] T. Ojala, M. Pietik¨ainen, and T. M¨aenp¨a¨a, “Multiresolution gray-scaleand rotation invariant texture classification with local binary patterns,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, 2002.

[15] N. Dalal and B. Triggs. Histograms of oriented gradientsfor human detection.In *Proc. CVPR*, 2005.

[16] Rakshit, S.; Anderson, C.H., "Computation of optical flow using basis functions," IEEE Transactions on Image Processing, vol.6, no.9, pp.1246,1254, 1997.

[17] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, “Evaluation of local spatio-temporal features for action recognition,” in *BMVC*, 2009, pp. 124–1.

[18] K. Ranganarayana and G. VenkateswaraRao,”Motion detection in low resolution video surveillance data to provide personal privacy”, in IJAER Vol.14 No.23 (2019) pp. 4251-4255.

[19] K. Ranganarayana and G. VenkateswaraRao,” Human recognition Using ‘PSO-OFA’ In Low Resolution Videos”, in Turkish Journal of Computer and Mathematics EducationVol.12 No.11 (2021), 697-703.

Authors Profile*:*

Katakam Ranga Narayana is perusing. PhD in IT department of GITAM University. His areas of interest Image processing

G.Venkateswara Rao*,* He is presently working as Professor in CSE department of GITAM University. Computer Networks and Mobile Computing,Cryptography and Network Security and Image Processing,.