STIP based Activity Recognition

Yashwanth K R
Students, Department of CSE,
NIEIT, Mysuru

Sunay M N
Students, Department of CSE,
NIEIT, Mysuru

Srinivas S
Students, Department of CSE,
NIEIT, Mysuru

Abhishek Rao
Students, Department of CSE,
NIEIT, Mysuru

Usha M S
Associate Professor, Department of CSE,
NIEIT, Mysuru

***Abstract*—The videos containing actions performed by persons were collected. The videos were converted into frames and the frames were preprocessed. Preprocessing is done through applying median filter. Features were extracted from the frames. Harris STIP, Gabor STIP and HOG STIP were used to extract the feature values from the video frames. The features were then classified using SVM classifier based on the kernel function. The classifier gives the label parallel to the input feature. The action is recognized based on the label returned by the classifier.**

***Keywords*—Action recognition, STIP(Spatio Temporal Interset Point), Harris filter, Gabor Filter. Histogram Orient Gradient (HOG)**

I. INTRODUCTION

Human action recognition employs a big role in human-to-human communication and mutual relations. As a result of it provides info regarding the identity of someone, their temperament, and condition, it's tough to extract. Recognition of actions within the video isn't a matter for human sensory system. The identification of the actions of the person by the system wants some special mechanisms. The recognition of the actions by the systems are going to be useful in computer vision method. This method is divided into low level action recognition process and high-level recognition process. Recognizing the actions victimization, the feature values extracted goes under low level action recognition method. These methods were simple to implement and that they don't seem to be reliable all the time. The high-level action recognition method needs some special hardware's to discover the actions within the video. These methods were a lot of reliable and that they were computationally expensive.

The videos containing actions performed by persons were collected. The videos were converted into frames and also the frames were pre-processed. Pre-processing is completed by applying median filter. The median filter finds the noises within the frame and replaces the noise by replace the element victimisation the median of the neighbor pixels. Options were extracted from the frames. The extracted options are going to be accustomed acknowledge the action of the person within the video. Harris SPIT, Dennis Gabor SPIT and HOG SPIT were used to extract the feature values from the video frames. The options were then classified victimisation SVM classifier supported the kernel operate. The action is recognized supported the label came back by the classifier. The performance of the planned technique is obtained by activity the performance of the classifier. The Accuracy, Sensitivity and Specificity of the classifier is that the most ordinarily lived performance measure for the classifier. This shows that our technique will considerably improve classification, performance of the video pictures.

1.1 Problem statement

Our work is to find out actions or events which are usually extremely unconstrained videos of day-to-day things. STIP-based sampling, local descriptors is also extracted on motion trajectories. The method using motion trajectories involves following and dense multi-scale optical flow computation, the associated machine complexity is higher. The real time videos contain noise.

1.2 Solution strategy

The videos containing actions performed by persons were collected. The videos were converted into frames and also the frames were pre-processed. Pre-processing is finished by using median filter. Features within the video is extracted using STIP based mostly approach. The options were then classified using SVM classifier based on the kernel operate. SVM trains the feature values based on the kernel functions and arranging the hyper-planes based on the labels of the classifier. The classifier provides the label like the input feature. The action is recognized based on the label came by the classifier. The performance is obtained by measure the performance of the classifier.

1.3 Action recognition

1.3 Classifications of HAR: Mainly there are two types of person action recognition, they are explained here

1.3 (a) Single User, Sensor-Based Action Recognition With machine learning and new data processing we develop a variety range of human actions by consolidating the rising space detector networks using the technique Sensor-based action recognition.

Mobile devices offer sufficient detector information and measuring power to allow accurate action recognition to get an energy consumption estimation of throughout day-to-day life. The sensor-based activity recognition researchers feel that computers are better desirable to act on our behalf to observe the behavior of agents.

1.3 (b) Multi-User, Sensor-Based Action Recognition

On-body sensors action was recognized for multiple users in early 90s. Throughout workplace scenarios detector technology like acceleration sensors, recognize the cluster activity patterns. With this, they question the basic problem of identifying actions of many users from detector measurements. Each single-user and multi-user actions in a

unified answer are know by proposing a novel pattern sound approach.

II. RELATED WORKS

The automation in every area comes by rapid growing of technology. The human face and facial expression recognition is heavily needed to specific applications in real life of person. It has many specific applications, which are data privacy, Image or video security surveillance, information security, biometric identification, Human Computer Interface (HCI), Human Behavior Interpretation (HBI), etc.,

As mentioned in first section, the human action recognition using STIP method ignores the spatial temporal (ST) inter relationships between the all types of person visual features. To improve the activity recognition there are many works have been presented to capture STIP information.

In the year 207-8, authors took the challenge in the field of leveraging vision of computer techniques in order to enrich HRI techniques, this concept explores the systems which can expand the capabilities of action. In the year 208, authors analyzed to detect and recognize activities using wearable sensor or mobile data which are collected with appropriate sensors. They have presented that feature extraction is very important stage in order to helps for reducing time of execution and improvement of accuracy of all person action. The authors Van and Tran, have proposed a techniques which exhibits both optical flow and RGB for HAR. They have analyzed the techniques and application of convolutional neural network, this CNN is very much suitable for the task of person visual activity recognition from various input videos .

A. Existing System

HAR mechanism provides description, interpretation, or comprehension of the scene by bringing out vital options from image. The flawless process can't be outlined as, recasting the present image in an exceedingly needed manner, and the output of the positioning action and speed is obtained at an equivalent time and real-time aspects. Innumerable SIFT variants were projected in order to spot the actions of the person in the video. SIFT-based sampling and local descriptors are often extracted on the motion trajectories.

B. Proposed System

The most well-built image processing system which consists of human eye together with the brain is the Human Visual system. With this resource we try to develop a computer vision system. The video is fed to the system which divides it into each different frames, preprocesses it, to reinforce the image frames by removing the unwanted pixels from the frames. By this technique we can reduce the noise and store those derived pictures for later usage. Options were extracted using the SIFT descriptors of various sort from the preprocessed video frames.

III. METHODOLOGY

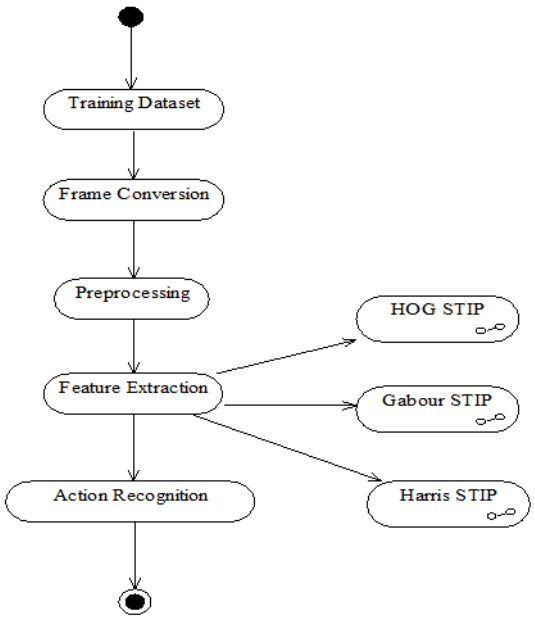

Fig. Action recognition flow diagram

The flow diagram of human action recognition is as shown in Figure. To get rid of the noise from the video, the input video frames were preprocessed. The performance of the method is improved by reducing the noise within the video. Many various forms of noise are present. In commonly noise is salt and pepper noise. In this occurs white and black pixels. The unwanted pixels are removed from the video frames by preprocessing. We will apply some filters to remove the noises from the video frames. The noised pixel in the image is detected by the median filter. The identified noisy pixel is replaced by the median value of the neighboring pixels. From preprocessed video frames the options were extracted using the STIP descriptors of various sort. The information carried in every of the thought-about descriptors is calculated. The un-normalized descriptors were extracted from the cuboids around STIP detections within the set of undistorted videos. The options like Harris STIP, Gabor STIP and HOG STIP were calculated and extracted from the frames. The Harris STIP is used to detect the corners within the video frames. This rule is used to detect corner in every pixels of the image, with the help of differential of the corner with reference to direction. Between the 2 patches there's the sum of squared differences (SSD). The similarity is indicated using the low numbers. If the nearby edges look similar,the it is said to be in uniform intensity. The Gabor STIP is used to identify the corners from the exact location of the object by using the Gabor wavelets. Local spectral energy density is provided by Gabor function. The convolution of two perpendicular directions is performed with variously dilated wavelets. The HOG STIP gives us the histogram values of the gradient at each point. The image is divided into small connected regions, called cells, and for each cell is complied with a histogram of gradient directions or edge detection for the pixels in the cell. The descriptor is indicated as the combination of these histograms. For improved accuracy, the local histogram can be contrast normalized by measuring the

intensity across the large region of the image called a block; these values are used to normalize all cells in the block. The person action is recognized in these videos. The performance is calculated by measuring the accuracy, sensitivity and specificity of the classifier.

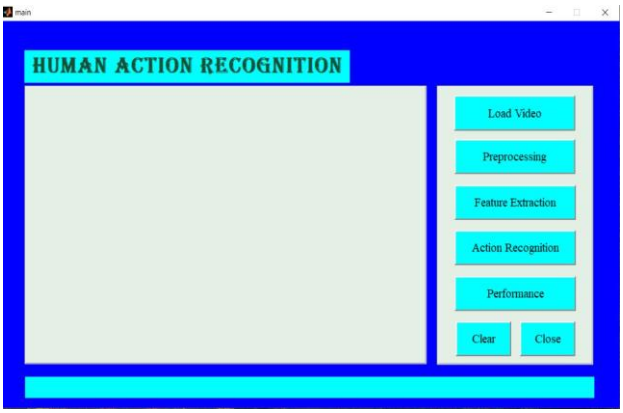

Fig.2. Human action recognition GUI window

IV. FEATURE EXTRACTION

A. *Harris STIP*

The algorithm is used to spot the corner present in each pixel of a picture using the corner score differentiation into account w.r.t direction. A grip is the sudden modification in the brightness of a picture. Corner is the junction of 2 edges. The resemblance is computed by locating the sum of squared differences (SSD) between the 2 patches. If the pixels within the image is of uniform intensity then the nearby edges will look similar if not the edges will look relatively different. To abstract some varieties of options and deduce the contents of a picture in computer vision systems Corner identification is a worthy appeal. Corner identification is applied many times in motion or movement's detection, image mosaicking, image registration, tracking of videos, panorama sewing, and 3D modeling and various types of objects recognition.

Detection process of Harris Corner

Intensity variation mechanism is used to detect all points through a local neighborhood we make use of Harris mechanism, and a very small region of the feature could be showing the maximum change in intensity levels when comparing with the shift of windows in any direction. This concept is explained using the autocorrelation functions are illustrated below:

Let us consider P as a scalar function which is represented by function $P \rightarrow R$ and small increment among any position in the domain as represented by h, $a \in \Omega$. Corners are defined as the points x that gives large values of the below illustrating functional for very small shifts h,

$$E(h) = \Sigma\, w(a)\,(P(a+h) - P(a) \quad (i)$$

That is the large variation in any other direction. The function w(a) gives permission for selecting the region of support, which is clearly called as a Gaussian function.

Taylor expansions will be used to get linearization of the expression P(a+q) as

$P(a+q) \simeq P(a)+\nabla(a)Tq$ Hence the right hand of (i) gives

$$E(q) \simeq \Sigma\, w(a)\,(\nabla P(a)\, q)2\, da =$$

$$\Sigma w(a)\,(qT \nabla P(a) \nabla P(a)\, Tq) \quad (ii)$$

The last equation (ii) depends on the image gradient through the matrix of autocorrelation, or tensor structure, which is represented as

$$Z = \Sigma\, w(a)\,(\nabla P(a) \nabla P(a)T) \quad (iii)$$

The largest eigen value of Z corresponds to maximum intensity variation direction, and also the second one corresponds to orthogonal direction of the intensity variation.

B. *Gabor STIP*

Local spectral energy density is provided by Gabor function. A two-dimensional convolution with a circular (non-elliptical) Gabor function is separable to series of one-dimensional ones. The convolution in 2 perpendicular directions is performed with multifariously expanded wavelets. It's necessary to use a ripple which is the first order partial differential operator. A Gabor wavelet which is the second order partial differential operator is used to discover the corners.

In image processing,it is a linear filter which is used for edge detection. Frequency and orientation representations of Gabor filters are like those of the human visual system, and that they are found to be particularly appropriate for texture illustration and discrimination. Thus, image analysis with Gabor filters is assumed to be like perception within the human visual system.

Features of Gabor filter: The basic feature extraction of Gabor filter in the two-dimensional function is as illustrated in expression.

The Gabor features referred to multiple resolution Gabor feature, are generated from outputs of Gabor filters by using multiple filters on many frequencies' fa and orientations. Frequency representations are illustrated in the equation

$a=h-a\ a= \{0,\ldots., A-\}$

Where, a is the a^{th} frequency, m = 0 is the max frequency generated and h> is the scaling factor of frequency. Let us consider as filter orientations are drawn as,

$= 2\ /N, = 0,\ldots., -2$

Where, is the *h* orientation and N is maximum orientations.

C. *HOG STIP*

HOG is known as Histogram of Oriented Gradients is the best object detection in image processing which uses the applications of feature descriptors. Fundamentally, the split of single image into very small connected regions which are called cells, and for each cell we compute a HOG directions. Each pixel of cell provides gradient weights to its respective angular bin. We can take blocks as spatial regions, which are the neighboring cells group. The base for classification and normalization of histograms is assembling of cells as blocks. The block diagram represents

the normalized group of histograms. This process yields better invariance to changes in brightness or shadowing.

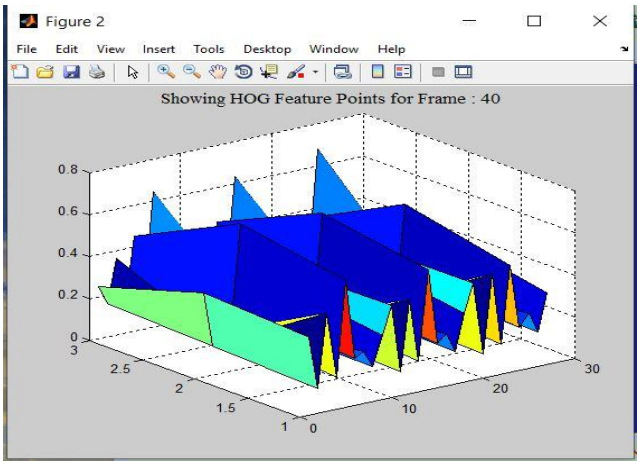

Fig.3. HOG feature points showcase

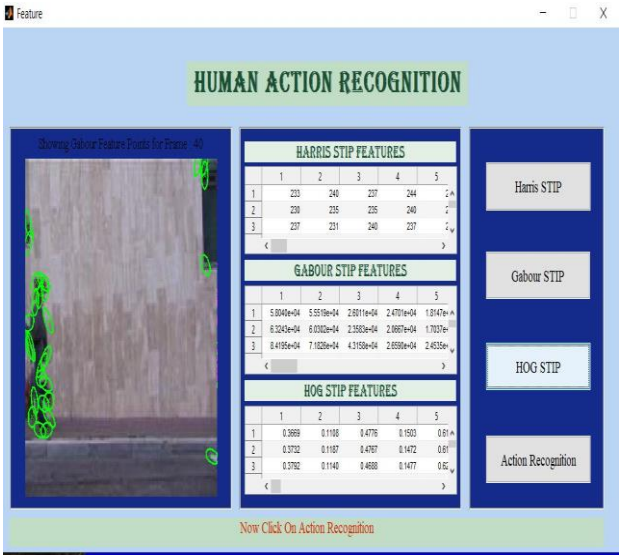

Fig.4. HAR feature extraction showcase

Calculation of Histogram Orient Gradient:

The initial step of generating the descriptor in HOG is to measure the one-dimensional derivatives point such as Ga and Gb in a and b direction by the convolution of gradient masks Ma and Mb with original image I:

$Ga = Ma * I$ $\quad Ma = [-\ 0\ ]$----------

$Gb = Mb * I$ $\quad Mb = [-\ 0\ ]^{T}$ ----------

With the help of derivatives basis functions Ga and Gb, which calculates the degree of HOG gradient [|G (a, b) |] and angle in direction ϕ (a, b) for each one of pixel.

The degree of HOG gradient shows its strength at a pixel is as shown in the equation 5:

$$|G(a,b)| = \sqrt{Ga(a,b)2 + Gb(a,b)2}$$----------5

The feature of 3D HOG extraction is as shown in Fig 4. All three feature extraction techniques in HAR system of STIP algorithm are shown in Fig 5.

sponsor acknowledgments in the unnumbered footnote on the first page.

V. ACTION RECOGNITION SYSTEM

Using Multi SVM classifier, the actions in the video are recognized. Support vector machines are utilized for the classification purpose. Support Vector Machine (SVM) is a non – probabilistic binary linear classifier. Expression for hyper plane is represented as (a.h)+t =0 Where, t – Set of training vectors, a – Vectors perpendicular to the separating hyper plane and h – Offset parameter which permits to raise the margin. The Output showing as “Running” and “Cycling” is one of the actions identified from the input video processing is illustrated in

$$\textbf{Sensitivity} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\textbf{Specificity} = \frac{True\ Negative}{False\ Positive + True\ Negative}$$

$$\textbf{Accuracy} = \frac{(True\ Positive\ (TP) + False\ Negative\ (FN))}{(False\ Positive\ (FP) + True\ Negative\ (TN)) + (True\ Positive\ (TP) + False\ Negative\ (FN))}$$

Fig.5. Illustration of performance measurements in HAR system.

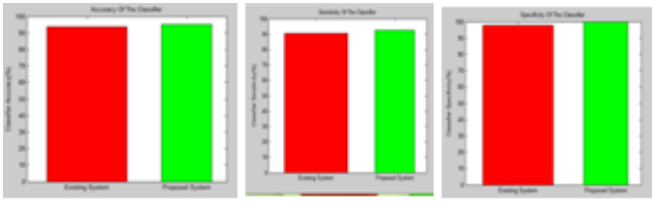

Fig. 6: Bar graph outputs for representation of Accuracy Sensitivity and Specificity of the given activity.

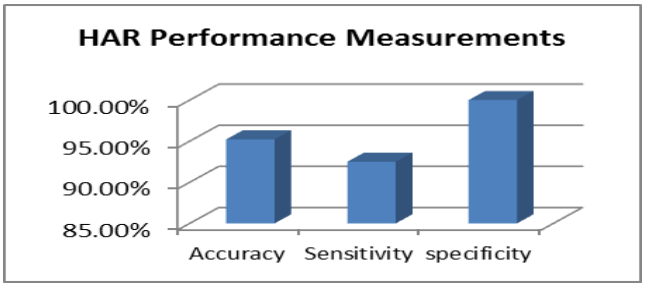

Fig.7. HAR Performance Measurements

VI. CONCLUSION

The planned system acknowledges the action of the persons within the video based on the options extracted using color stips. The extracted features area unit based on the STIP that is combined with several of the opposite ideas so the feature extraction method is simpler. The recognition of the action is completed using the kernel function of the SVM classifier. The planned system provides accuracy that is more than the present algorithms that identifies that the misclassifications area unit reduced to a larger extend. The improved modeling of appearance results in an improved balance between photometric invariance and discriminative power, as chromaticity provides a lot of info, based on that higher representations are formed. Color stips are totally evaluated and shown to considerably beat out their intensity-based counterparts for recognizing human actions on variety of difficult video benchmarks. The planned method recognizes the action of the person within the video supported the options extracted exactly even if there have been numerous challenges like illumination variations, contrast variations, Abrupt motions and Scaling of the persons in the video.

VII. FUTURE WORK

The human action recognition techniques can also be applied using data fusion techniques. Which are speech action data can add to the human facial expressions or any action recognition to achieve better performance output to the real time given videos.

VIII. ACKNOWLEDGMENT

This project was completed under the guidance of Smt. USHA M S, Associate Professor, NIE Institute of Technology, Mysore. The Authors would like to extend their heartfelt gratitude to Smt. USHA M S and all the people who are directly or indirectly involved, advised and coordinated in collecting data and software information for the support of this paper.

REFERENCES

[1] Md. Atiqur Rahman Ahad, J.K. Tan, H.S. Kim and S. Ishikawa, “Human Activity Recognition: Various Paradigms”, International Conference on Control, Automation and Systems 2008.

[2] S. C. F. Lin, C. Y. Wong, T. R. Ren and N. M. Kwok, “A Comparison Study on Human Action Recognition from Video Streams”, 978--4673-0964-6/2, 202 IEEE.

[3] Ong Chin Ann, Lau Bee Theng, “Human Activity Recognition: A Review”, 978--4799-5686-9/4, 204 IEEE.

[4] Xue Li, Quan Z. Sheng, ChaoyiPang , Xin Zhao, and Sen Wang, “Effective Approaches in Human Action Recognition”, ISBN: 978-979-42-9-5, 203 IEEE.

[5] Luo Dan, Hazım Kemal Ekenel, Ohya Jun, “Human Gesture Analysis using Multimodal features”, 202 IEEE International Conference on Multimedia and Expo Workshops.

[6] Chen Chen • Kui Liu • Nasser Kehtarnavaz, “Real-time human action recognition based on depth motion maps”, 203 Springer.

[7] Suraj Vantigodi, R. Venkatesh Babu, “Real-time Human Action Recognition from Motion Capture Data”, 204 IEEE.

[8] Nilam Nur Amir Sjarif, Siti Mariyam Shamsuddin, “Human Action Invarianceness for Human Action Recognition”, 978--4673-6744-8/5, 205 IEEE.

[9] Neziha JAOUEDI, Noureddine BOUJNAH, Oumayma HTIWICH, Med Salim BOUHLEL, “Human Action Recognition to Human Behavior Analysis”, 978--5090-472-3/6, 206 IEEE.

[10] Chandni J. Dhamsania, Prof. Tushar V. Ratanpara, “A Survey on Human Action Recognition from Videos”, 978--5090-4556-3/6, 206 IEEE.

[11] Alexandre Perez, Hedi Tabia, David Declercq and Alain Zanotti, “Feature covariance for human action recognition”, 978--4673-890-5/6, 206 IEEE.

[12] [12] Di Wu, Nabin Sharma, and Michael Blumenstein, “Recent Advances in Video-Based Human Action Recognition using Deep Learning: A Review”, 978--5090-682-2/7, 207 IEEE.

[13] Pravin Dhulekar, S.T. Gandhe, Harshada Chitte and Komal Pardeshi, “Human Action Recognition: An Overview”, Springer 207.

[14] Soumalya Sen, Moloy Dhar, Susrut Banerjee, "Implementation of Human Action Recognition using Image Parsing Techniques", IEEE 2009.

[15] Francesco Monti, Carlo S. Regazzoni, "Human Action Recognition using the motion of Interest Points", 200 IEEE 7th International Conference on Image Processing.

[16] Manuel J. Marin-Jimenez, Enrique Yeguas, Nicolas Perez de la Blanca, "Exploring STIP-based models for recognizing human interactions in TV videos", 202 Elsevier.

[17] Ivo Everts, Jan C. van Gemert, and Theo Gevers, "Evaluation of Color Spatio-Temporal Interest Points for Human Action Recognition", IEEE Transactions on Image Processing, Vol. 23, No. 4, April 204.

[18] Guillermo C´amara-Ch´avez, Arnaldo de Albuquerque Ara´ujo, "Harris-SIFT Descriptor for Video Event Detection based on a Machine Learning Approach", 2009 th IEEE International Symposium on Multimedia.

[19] Feng Zhu, Xian-Da Zhang, Ya-Feng Hu, "Gabor Filter Approach to Joint Feature Extraction and Target Recognition", IEEE Transactions on Aerospace And Electronic Systems Vol. 45, No. January 2009.

[20] Yuanyuan Huang, Haomiao Yang, Ping Huang, "Action Recognition Using HOG Feature in Different Resolution Video Sequences", 978-0-7695-4639-/2, 202 IEEE.

[21] Mohamed Ibn Khedher, Mounim A. El-Yacoubi and Bernadette Dorizzi, "Human Action Recognition using Continuous Hmms and HOG/HOF Silhouette Representation", ICPRAM 202.

[22] Huimin Qian , Yaobin Mao, Wenbo Xiang, Zhiquan Wang, "Recognition of human activities using SVM multi-class classifier", 202 Elsevier.

[23] Adithyan Palaniappan, R. Bhargavi, V. Vaidehi, "Abnormal Human Activity Recognition Using SVM Based Approach", ISBN: 978--4673-60-9/2, 202 IEEE.

[24] Fam Boon Lung, and Mohamed Hisham Jaward "Spatio-Temporal Descriptor for Abnormal Human Activity Detection", 978-4-9022-4-6, 4th IAPR International Conference on Machine Vision Applications (MVA), May 8-22, 205.

[25] Debapratim Das Dawn, Soharab Hossain Shaikh, "A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector", Springer 205.

[26] Tehmina Kalsum, Syed Muhammad Anwar, Muhammad Majid2, Bilal Khan3, Sahibzada Muhammad Ali, "Emotion recognition from facial expressions using hybrid feature descriptors", IET Image Process., 208, Vol. 2 Iss. 6, pp. 004-02.

[27] Parameshachari B D et. al "Predictive Patch Matching Method For Inter Frame Coding In Advanced Video Coding", 2017 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), PP 918-922, 978-1-5386-2361-9/17/$31.00 ©2017 IEEE.

[28] Prabu, S., M. Lakshmanan, and V. Noor Mohammed. "A multimodal authentication for biometric recognition system using intelligent hybrid fusion techniques." Journal of medical systems 43.8 (2019): 249.