Grenze International Journal of Engineering and Technology, June Issue

Development of HAR using Spatio Temporal Machine Learning Techniques

Suchithra N P[1] and K Viswanath[2]
[1]Research Scholar, Department of Telecommunication Engineering, Siddaganga Institute of Technology, Tumakuru, India,
Email: suchithra.mtech@gmail.com
[2]Professor, Department of Telecommunication Engineering, Siddaganga Institute of Technology, Tumakuru, India
Email: viswasit@gmail.com

***Abstract*— Action recognition of human aid us to spot their action their nature, and emotional state, that is hard for extraction. Hence, it plays a decisive role in person's interaction and communal relations. One of the primary subjects of study in the scientific fields of computer vision and machine learning is the human ability to recognize another person's activity. By the results of this analysis, several activity recognition system is needed for several applications, together with video surveillance systems, human-computer interaction, and robotics for human behavior characterization. For the analysis of image and video, human activity recognition (HAR) is an important analysis direction. In this paper, we offer an embrace survey of the recent development techniques, including strategies, systems, and quantitative analysis of the performance of HAR. By the experimental results it shows that our technique will remarkably improve classification, interpretation, and retrieval performance for the video pictures. The novelty of this paper is bipartite. Firstly, the video pictures of human is captured. Secondly, different types of actions performed by humans are going to be identified.**

***Index Terms*— Action recognition, Spatial-Temporal Interest Point (STIP), Gabor Filter.**

I. INTRODUCTION

The primary objective of HAR is to automatically identify multiple activities from given videos. The action recognized from the video is vital for Video categorization, retrieval and security purpose. Human visual system isn't a matter for recognition of actions in the video. The recognition of human activities is one of the important issues in surveillance video [1] – [3]. However, some special mechanisms are needed to identify the actions of human by the computer system. The action recognized by the systems are employed in computer vision method. This method is often divided into two levels: low level recognition process and high level recognition process. Low level action recognition method comes under recognizing the actions from the extracted feature values. It is simple to implement, but not reliable all the time. The high level action recognition method is employed to find actions in the video using some special hardware. It is more reliable, but it is computationally expensive. The main goal of action recognition is to identify the actions of one or more person from a continual observation on the person's actions and changes within the environmental conditions. Because of its multi-aspect in nature, action recognition system is often referred in several fields as arrange recognition, goal recognition, intent recognition, behavior recognition, location estimation and location-based services.

Grenze ID: 01.GIJET.12.2.510_1

Ivo Everts et al. [4] proposed to automatically select from a pool of descriptors, which is best suitable based on object surface and scene properties.

A. Existing system

The machine provides description, interpretation, or understanding of the scene by extracting vital options from image. Correct process can't be outlined as, altering a present image in an exceedingly needed manner, and at an equivalent time the real-time output of the positioning action and speed is obtained. To identify the actions of the persons in the video different SIFT variants were projected. On the motion trajectories, Spatial-Temporal Interest Point (STIP)-based sampling and local descriptors are often extracted.
An effective method to detect certain human actions even within crowd sequences of surveillance videos is proposed by Masaki Takahashi et al. [5]. Manoranjan Paul et al. presented a comprehensive review and comparison of available techniques for detecting humans in surveillance videos in [6]. Action recognition methodologies are especially needed for surveillance systems that aim to prevent crimes and malicious activities before they occur. 3D – Convolutional Neural Networks (3D-CNN) with 3D motion cuboid for action detection and recognizing in videos. is proposed by J. Arunnehru et al. in [7]. Marek Kulbacki in [8] proposed another comprehensive overview of intelligent video analytics and human action recognition methods. Monji Mohamed Zaidi et al. proposed [9] to identify six suspicious acts (Running, Punching, Falling, Snatching, Kicking, and Shooting) that have not been previously examined in the existing literature. Kishore K. Reddy et al. [10] proposed the system which recognize 50 human action categories of web videos. Human actions in videos are recognized based on spatio-temporal interest points (STIPs) by [11], and human pose estimations by Michael B. Holte et al. [12]. Human motion detection for real-time security [13], identification of human facial expression using spatial temporal algorithm [14], and human identification based on Iris recognition using support vector machine [15] are proposed.

B. Proposed system

Human visual system is that the most powerful image process system that consists of human eye together with the brain. Keeping this truth in mind we try to develop a computer vision system. The System receives an input video, divides them into each different frames, preprocesses it, to reinforce the image frames by removing the unwanted pixels from the frames therefore reducing the noise and stores those derived pictures for later usage. Options were extracted from the preprocessed video frames using the STIP descriptors of various sort.

C. Types of Action Recognition

Sensor-based, single-user action recognition
Sensor-based action recognition is carried out to develop a wide range of human actions by consolidating the rising space of detector networks with novel data processing and machine learning techniques. To obtain an estimation of the energy consumption throughout day-to-day life, mobile devices offer sufficient detector information and calculation power to allow physical action recognition.

Sensor-based, multi-user action recognition
In the early 90s, using on-body sensors action was recognized for multiple users, 1st arose in the work by ORL using the active badge systems. Throughout workplace scenarios cluster activity patterns were recognized by detector technology like acceleration sensors. By this work, they interrogate the basic problem of identifying actions of multiple users from detector readings in a home environment and each single-user and multi-user actions in a unified answer are known by proposing a novel pattern sound approach

Vision-based action recognition
Vision based human action recognition is that, the method of labeling image sequences with action labels. The videos taken by numerous cameras facilitate us to trace, and understand the behavior of agents that are very important and difficult problem. The applications of vision-based action recognition include human-computer interaction, interface design, robot learning, and surveillance, among others.

II. HUMAN ACTIVITY RECOGNITION (HAR)

The flow diagram of human activity recognition is as shown in the Figure 1. The noise from the video shall be removed by the input video frames using pre-processing. The performance of the process is improved by reducing the noise in the video. Many various varieties of noise are present in frames images. The usually obtained noise is salt and pepper noise. It presents lightly occurs white and black pixels. The unwanted pixels

are far from the video frames by pre-processing. Noise from the video frames can be removed by applying any of the filters available. The noised pixel in the image is detected by the median filter. The known noisy pixel is replaced by the average of the neighbouring pixels. From pre-processed video frames the options were extracted using the STIP descriptors of various kind.

The information carried in each of the considered descriptors is calculated. The un-normalized descriptors were extracted from the cuboids around STIP detections in the set of undistorted videos. The features like Harris STIP, Gabor STIP and HOG STIP were calculated and extracted from the frames. The Harris STIP is used to detect the corners in the video frames. This algorithm is used to detect corner in each pixels of the image, by considering the differential of the corner with to direction. Between the two patches there is the sum of squared differences (SSD). The process of input video loading is as shown in the Figure 2, which clearly show case of the size of the input video and frames counts per second for image analysis.

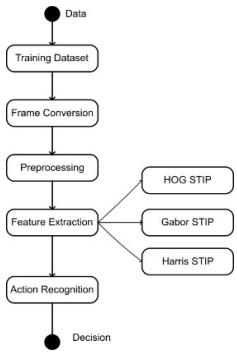

Figure 1. Flow Diagram of human activity recognition (HAR)

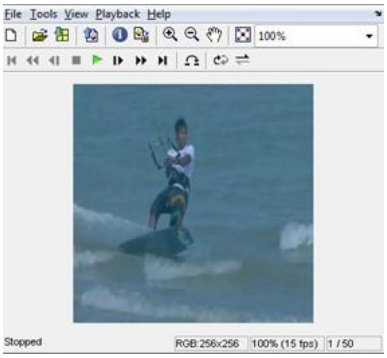

Figure 2. Loading the input video to the HAR System

A lower number indicates more similarity. If the pixel is in a region of uniform intensity, then the nearby edges will lock similar. The Gabor STIP is used to identify the corners from the exact location of the object by using the Gabor wavelets. For a given position and frequency in a certain direction, the Gabor function provides a local spectral energy density. 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 video. The performance is calculated by measuring the accuracy, sensitivity and specificity of the classifier. After the loading input video in the HAR GUI window need to check for preprocessing of the video which is shown in the Figure 3. The frames and its counts of the given input video to the HAR System are illustrated in the Figure 4. These frames are very much useful for analyzing motion of an action in the STEP analysis.

III. Feature Extraction Techniques

Feature extraction techniques refer to the computational methods that transform raw data into a set of informative and non-redundant features suitable for analysis. Different types of techniques are explained subsequently.

Figure 3. Human Activity Recognition GUI Window representation

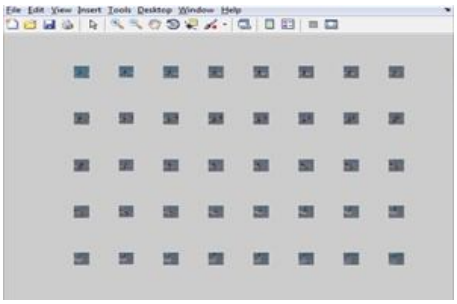

Figure 4. Illustration of converting video in to the frames

A. Harris STIP

Harris stips are local maxima of the 3D Harris energy function based on the structure tensor. The algorithm is used to spot the corner present in each pixel of a picture using the differential of the corner score into account with reference to direction. The sudden modification in the brightness of a picture is a grip. The junction of 2 edges is known as corner. The resemblance is calculated by considering the sum of squared differences (SSD) between the 2 patches. The nearby edges will look similar if the pixel within the image is of uniform intensity if not the edges will look quite different.

Detection process of Harris Corner

The Harris mechanism is used to detect all points based on the intensity variation of a local neighborhood, and a small region of the feature should show the large change in intensity when compared with windows shifted in any direction. This concept is expressed through the autocorrelation function as follows: Let the image be a scalar function I : → R and h a small increment around any position in the domain, x ∈ Ω . Corners are defined as the points x that maximize the following functional for small shifts h,

$$E(h) = \sum w\,(x)\,(I(x+h) - I(x)\,)^2 \qquad (1)$$

i.e. the maximum variation in any direction. The function w(x) allows selecting the support region that is typically defined as a Rectangular or Gaussian function. Taylor expansions can be used to linearize the expression I(x + h) as I(x + h) ≃ I(x) + ΔI(x)T h, so that the right hand of (1) becomes (2).

$$E(h) \cong \sum w\,(x)\,(I(x)h)\,)^2\,dx = \sum w(x)(h\,T\,\Delta I(x)\Delta I(x)Th \qquad (2)$$

This last expression depends on the gradient of the image through the autocorrelation matrix, or structure tensor, which is given by (3).

$$M = \sum w(x)(\Delta I(x)\,\Delta I)x\,T \qquad (3)$$

The largest eigen value of M corresponds to the direction of largest intensity variation, while the second one corresponds to the intensity variation in its orthogonal direction.

B. Gabor STIP

A local spectral energy density concentrated around a given position and frequency in a certain direction is provided by Gabor function. A 2D convolution with a circular Gabor function is separable to series of one-dimensional ones. A Gabor wavelet which serves as the second order partial differential operator can be used to detect the corners. Gabor filter is a linear filter for detecting edges and it is named after Dennis Gabor. Frequency and orientation description of Gabor filters are same as that of human visual system, and they are found to be peculiarly suitable for texture description and differentiation. In spatial domain, a 2D Gabor filter is a Gaussian Kernel Function regulated by a sinusoidal plane wave.

Gabor Features: A core of Gabor filter based feature extraction is the 2D Gabor filter function expressed as

$$\Psi(x,y) = f\text{^}2/(\pi\,\gamma\,\eta)\cdot exp(-(f\text{^}2/\gamma\text{^}2\cdot x'\text{^}2 + f\text{^}2/\eta\text{^}2\cdot y'\text{^}2))\cdot exp(-j\cdot 2\pi\cdot f\cdot x'). \qquad (4)$$

In the spatial domain the Gabor filter is a complex plane wave (a 2D Fourier basis function) multiplied by an origin-centered Gaussian. f is the central frequency of the filter, θ the rotation angle, γ sharpness (bandwidth) along the Gaussian major axis, and η sharpness along the minor axis (perpendicular to the wave). In the given form, the aspect ratio of the Gaussian $\eta/\ \gamma$. Gabor features, referred to as Gabor jet, Gabor bank or multiresolution Gabor feature, are constructed from responses of Gabor filters in (1) or (2) by using multiple filters on several frequencies f_m and orientations θ_n. Frequency in this case corresponds to scale information and is thus drawn from equation (5).

$$f_m = k - m\,f_{max}\,, m = 0, \dots, M-1 \qquad (5)$$

where, f_m is the m^{th} frequency, f_{max} is the highest frequency desired and k>1, is the frequency scaling factor. The filter orientations are drawn as in the equation (6).

$$\theta_n = \frac{2\,\pi\,n}{N}\,, n = 0, 1, \dots, N-1 \qquad (6)$$

where θ_n is the N^{th} orientation and N is the total number of orientations. Scaled of a filter bank are selected from exponential (octave) spacing and orientations from linear spacing.

C. HOG STIP

The object detection in computer vision and image processing make use of a feature descriptors known as Histogram of Oriented Gradients (HOG). HOG is widely used as a feature described image region for object detection such as human face or human body detection. In this process first divide the image into small connected regions called cells, and for each cell compute a histogram of gradient directions or edge orientations for the pixels within the cell. Each cell's pixel contributes weighted gradient to its corresponding angular bin. Groups of neighboring cells are taken as spatial regions called blocks. The assembling of cells as blocks is the core for classification and normalization of histograms. Normalized group of histograms represents the block histogram. The descriptor shows the group of these block histograms. This normalization provides better invariance to changes in brightness or shadowing.

Calculation of Histogram Orient Gradient:

The first step for producing the HOG descriptor is to calculate the 1-D point derivatives G_x and G_y in x- and y direction by convolving the gradient masks M_x and M_y with the raw image I:

$$G_x = M_x * I, M_x = [-101] \quad (7)$$

$$G_y = M_y * I, M_y = [-101]^T \quad (8)$$

On the basis of the derivatives G_x and G_y we then calculate the gradient degree G(x, y) and direction angle ϕ (x,y) for each pixel. The gradient degree shows the gradient strength at a pixel is given by (9).

$$|G_{x,y}| = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \quad (9)$$

We do not omit the extraction of the square root since the performance study of Dalal and Triggs describes best results with this Euclidean metric [1] The feature of HOG extraction is as shown in Figure 5. All three feature extraction techniques in HAR system of STIP algorithm are shown in Figure 6.

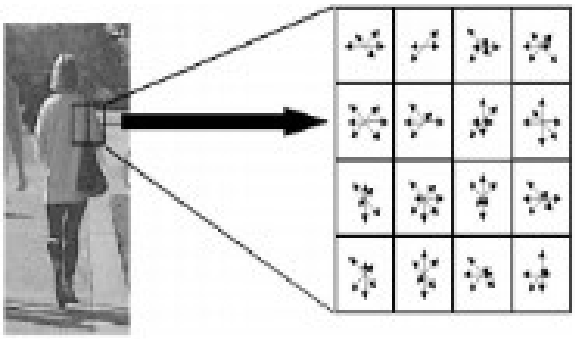

Figure 5: Extraction Process of HOG features

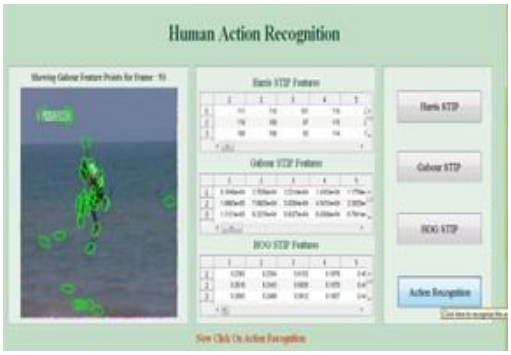

Figure 6: Illustration of feature extraction techniques in HAR System

IV. ACTION RECOGNITION SYSTEM

The actions in the video are recognized using Multi SVM classifier. Support vector machines that are used for classification and regression analysis, are supervised learning models with associated learning algorithms that can analyze data and recognize patterns. SVM is a non-probabilistic binary linear classifier. Expression for hyper plane is $w \cdot x + b = 0$, where x is set of training vectors, w is vector perpendicular to the separating hyper plane, and b is the offset parameter which permits to raise the margin.

Kernel function can be used while decision function is not a linear function of the data and the data can be plotted from the input space from a nonlinear conversion instead of placing non-linear curves to the vector space for separating the data. The classification task is now able to scale high dimensional data relatively well, as an optimal kernel function is implemented in SVM model, Therefore, the tradeoff between classifier complexity and classification error can be controlled explicitly. In machine learning, support vector machines are used as supervised learning models with associated learning algorithms that used for classification by data analysis and pattern recognition. By using a set of training examples, each noticed as belonging to one of two classifications, an SVM training algorithm constructs a model that allocate new examples into one classification or the other. In addition to carry out linear classification, SVMs can precisely perform a non-linear classification by using the kernel trick, completely plotting their inputs into high-dimensional feature spaces.

Figure 7 shows the plots used by SVM schemes are sketched to protect that dot products can be determined easily in terms of the variables in the primary space, by explaining them in terms of a kernel function K (x, y) choose to suit the complication. The hyperplanes in the higher-dimensional space are explained as the group of points whose dot product with a vector in that space is idle. The vectors that define the hyperplanes are chosen to be linear combinations with parameters αi of images of feature vectors that appear in the data base. By taking the hyperplanes, the points in the feature space that are plotted to the hyperplane are explained from the relation in (10). Note that if K (x, y) becomes little as y grows more away from x, each term in the sum calculates the degree of closeness of the test point x to the correlative data primary point xi. In this way, the sum of kernels above can be used to detect the parallel nearness of each test point to the data points. The set of points x plotted into any hyperplane can be fully convoluted. As a result, much more complex discrimination between sets are allowed which are not convex at all in the original space. The Output showing as "Surfing" is one of the action identified from the input video processing is illustrated in Figure 8.

$$\sum_i \alpha_i K(x_i, x) = constant \qquad (10)$$

V. Performance Measures

The accuracy, sensitivity and specificity of the classifier are calculated to measure the performance of the system. The rate at which classifier identifies the image based on the given label represents the accuracy of the classifier. Sensitivity is even called as the true positive rate, or the recall rate in some discipline. The sensitivity of the classifier is calculated based on how exactly the classifier correctly classifies the data to each category. Specificity sometimes called the true negative rate. The specificity of the classifier is calculated based on how exactly the classifier rejects the data to each category. The following equations explains the calculation of Sensitivity, Specificity and Accuracy of the given input action video.

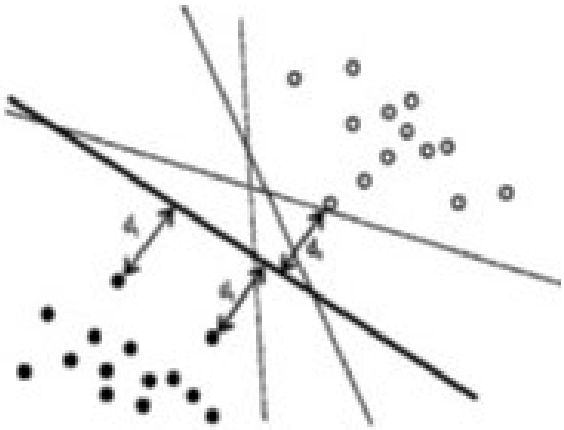

Figure 7: Illustration of a kernel functions in SVM Model

Figure 8: Output showing as "Surfing" is one of the action identified from the input video processing

$$Sensitivity = \frac{True\ Positive}{TruePositive + FalseNegative}$$

$$Specificity = \frac{True\ Negative}{False\ Positive + True\ Negative} \quad (11)$$

$$Accuracy = \frac{True\ Positive + False\ Negative}{(False\ Positive + True\ Negative)\ (True\ Positive + False\ Negative)} \quad (12)$$

The performance measurements in HAR system and its calculated results are illustrated in Figure 9. The Confusion matrix result of "Surfing" action in HAR system is as shown in Figure 10. The accuracy of the given existing and proposed system is as shown in Fig. 11 (a). The sensitivity of the given existing and proposed system is as shown in Fig 11 (b). The specificity of the given existing and proposed system is as shown in Fig. 11 (c). The ROC output of "Surfing" action in Human Action Recognition system is as shown in the Figure 11 (d).

Figure 9: Illustration of performance measurements in HAR system

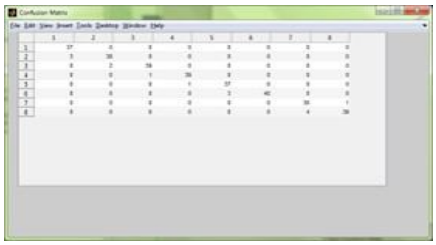

Figure 10: Confusion matrix result of "Surfing" action in HAR system

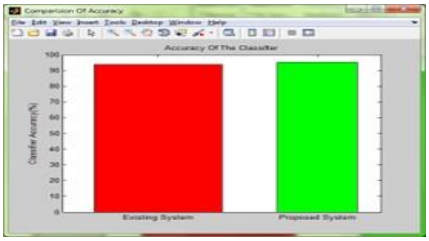

Figure 11 (a): Comparison of Accuracy of the given activity

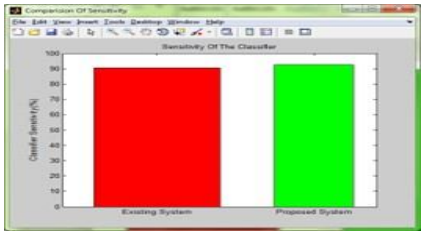

Figure 11 (b): Comparison of Sensitivity of the given activity

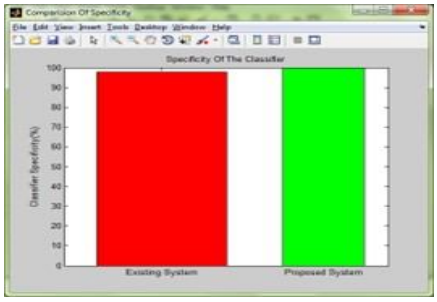

Figure 11 (c): Comparison of Specificity of the given activity

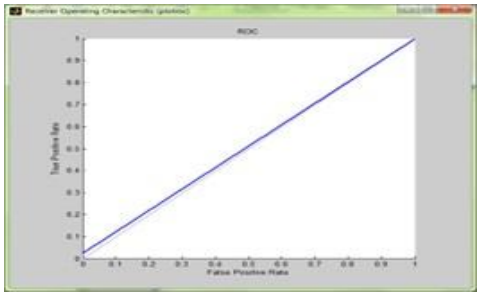

Figure 11 (d): ROC output of "Surfing" action in HAR

V. Conclusion and Future Enhancement

The planned system recognizes the action done by the person in the video based on the options extracted exploitation color STIPs. The options are extracted supported the STIP with the combination of the many different ideas specified the feature extraction method become more effective. The action recognition is created using the kernel function of the SVM classifier. The accuracy given by the planned system is higher than the existing algorithms during which the misclassifications are reduced to a larger extend. STIP detectors and descriptors were reformulated so that multiple photometric channels are incorporated additionally with image intensities, leading to color STIPs. Balance between photometric in variance and discriminative power is improved leading to higher representations and increased modelling of appearance. Color STIPs are evaluated

completely and their intensity-based counterparts will considerably outperform for recognizing human actions on variety of difficult video benchmarks.

The projected technique exactly recognizes the action of the person in the video based on the extracted options although there have been various challenges like illumination variations, contrast variations, abrupt motions and scaling of the persons within the video. The planned system isn't a completely automated system. So as to boost the process of the system, it is often automated by avoiding a supervised classifier. The supervised classifiers need manual labels and therefore the training of the system for the classification purpose. The system is often improved by as well as some further feature extraction algorithms which may describe a lot of information about the classifier. The additional options extracted ought to be able to overcome the problem of real time implementation of the process.

REFERENCES

[1] Ullah, W.; Ullah, A.; Hussain, T.; Khan, Z.A.; Baik, S.W. An Efficient Anomaly Recognition Framework Using an Attention Residual LSTM in Surveillance Videos. Sensors 2021, 21, 2811. https://doi.org/10.3390/s21082811

[2] Amin Ullah, Khan Muhammad, Weiping Ding, Vasile Palade, Ijaz Ul Haq, Sung Wook Baik, Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications, Applied Soft Computing, Volume 103, 2021, 107102, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2021.107102.

[3] Waseem Ullah, Amin Ullah, Tanveer Hussain, Khan Muhammad, Ali Asghar Heidari, Javier Del Ser, Sung Wook Baik, Victor Hugo C. De Albuquerque, Artificial Intelligence of Things-assisted two-stream neural network for anomaly detection in surveillance Big Video Data, Future Generation Computer Systems, Volume 129, 2022, Pages 286-297, ISSN 0167-739X, https://doi.org/10.1016/j.future.2021.10.033.

[4] Everts, I., van Gemert, J.C., Gevers, T. (2012). Per-patch Descriptor Selection Using Surface and Scene Properties. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds) Computer Vision ECCV 2012. ECCV 2012. Lecture Notes in Computer Science, vol 7577. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33783-3-13

[5] M. Takahashi, M. Naemura, M. Fujii and S. Satoh, "Human action recognition in crowded surveillance video sequences by using features taken from key-point trajectories," CVPR 2011 WORKSHOPS, Colorado Springs, CO, USA, 2011, pp. 9-16, doi: 10.1109/CVPRW.2011.5981713.

[6] Paul, M., Haque, S.M.E. and Chakraborty, S. Human detection in surveillance videos and its applications - a review. EURASIP J. Adv. Signal Process. 2013, 176 (2013). https://doi.org/10.1186/1687-6180-2013-176

[7] J. Arunnehru, G. Chamundeeswari, S. Prasanna Bharathi, Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos, Procedia Computer Science, Volume 133, 2018, Pages 471-477, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.07.059.

[8] Kulbacki, M.; Segen, J.; Chaczko, Z.; Rozenblit, J.W.; Kulbacki, M.; Klempous, R.; Wojciechowski, K. Intelligent Video Analytics for Human Action Recognition: The State of Knowledge. Sensors 2023, 23, 4258. https://doi.org/10.3390/s23094258

[9] M. Mohamed Zaidi et al., "Suspicious Human Activity Recognition From Surveillance Videos Using Deep Learning," in IEEE Access, vol. 12, pp. 105497-105510, 2024, doi: 10.1109/ACCESS.2024.3436653.

[10] Reddy, K.K., Shah, M. Recognizing 50 human action categories of web videos. Machine Vision and Applications 24, 971–981 (2013). https://doi.org/10.1007/s00138-012-0450-4

[11] I. Everts, J. C. van Gemert and T. Gevers, "Evaluation of Color STIPs for Human Action Recognition," 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 2013, pp. 2850-2857, doi: 10.1109/CVPR.2013.367.

[12] M. B. Holte, C. Tran, M. M. Trivedi and T. B. Moeslund, "Human Pose Estimation and Activity Recognition From Multi-View Videos: Comparative Explorations of Recent Developments," in IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 5, pp. 538-552, Sept. 2012, doi: 10.1109/JSTSP.2012.2196975.

[13] Pavithra S, Mahanthesh U, Stafford Michahial, Dr. M Shivakumar, "Human Motion Detection and Tracking for Real-Time Security System", International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 5, Issue 12, December 2016.

[14] Mahanthesh U,Dr. H S Mohana "Identification of Human Facial Expression Signal Classification Using Spatial Temporal Algorithm" International Journal of Engineering Research in Electrical and Electronic Engineering (IJEREEE) Vol 2, Issue 5, May 2016.

[15] Lalitha. K, Deepika T V, Sowjanya M N, Stafford Michahial, "Human Identification Based On Iris Recognition Using Support Vector Machines", International Journal of Engineering Research in Electrical and Electronic Engineering (IJEREEE) Vol 2, Issue 5, May 2016.