Extraction of posture silhouettes using human posture feature points and spatial masks for activity recognition in public place

Monika* and Madhulika Bhatia**

* Department of Computer Science & Engineering, Apex Institute of Technology, Chandigarh University, Punjab, 140413, India
** Department of Computer Science & Engineering, Amity School of Engineering and Technology, Amity University, Noida, 201301, India
** Corresponding Author: madhulikabhatia@gmail.com

ABSTRACT

Presently, an automated system has been required for public place security. Recognizing human postures in public places has emerged as a global solution for understanding human behavior in public places. In this work, a model to extract a human feature attribute of its posture has been presented to identify human behavior. The research work in this paper focuses on identifying the seating and standing postures of a person. The proposed methodology aims towards extraction of the human attributes from public places using spatial masks. Consequently, in this process, unwanted details from the background have been removed using the technique to focus on human postures only. The feature extraction process gives us blob vector and posture vector to evaluate human authentication and posture apprehension.

Keywords: Human body postures; Posture classification; Posture estimation; Head pose estimation; Behavior analysis.

INTRODUCTION

Enhancing surveillance system is an active topic in research. Human detection and behavior apprehension is the most significant domain for improving the research in this field. Human detection and behavior apprehension is a process to employ experts to prevent and report misdemeanor, suspicious activities, violence, and sabotage. Human detection and behavior apprehension contrivances can be used by experts for the avoidance of calamity or act in response to the investigation. A detailed study on the importance and improvement areas of the surveillance system is presented in (Monika and Madhulika, 2014). In Kadamvari et al., 2015, to improve surveillance system role of traffic lights, simulation methodologies were identified. The detection method includes differentiating human from the environment, that is, human versus non-human (Chen, Liang-Hua, Li-Yun Wang, and Chih-Wen Su, 2014) (where environments refer to airports, railways, buses, traffic signals, etc.). In contrast, apprehension methods include the single person (e.g., lurking) and multiple person interfaces such as public places like museums (Wang Z. and Nagai Y., 2018). In the past few years, the significance of human detection and behavior apprehension research

emerged enormously. It includes research on a human in the environment, their activities, and realization facts. A high-level computer vision technique needs to be modeled for optimal human behavior apprehension. Conventional methods for detection and tracking human are broadly categorized into two main approaches, context-based methods and foreground-based methods, in which tracking is recommended as a preparatory task. According to the latest record, the transit method for the Delhi metro in India needs approximately a hundred workers in support of routine procedures (DMRC).

Similarly, if additional public sections are combined, probably considerable numbers of CCTV cameras are required for producing a human-dependent surveillance system. It has been observed that analysis of the footage is not operationally feasible as the present system requires a large number of human analysts. But the new thwart emerged as large numbers of videos are challenging to store, capture, and share. So, the system needs to be more intelligent to analyze and prevent such causes.

This work proposes a feature extraction technique that contains minimum attributes to identify human behavior in public places. To affirm the vectors correlation, minimum and maximum range identification measure is used. The feature extraction process generates blob and posture vectors. The blob vector has been developed using the human body golden ratio rule of the posture vector. Blob and posture vectors have been used to detect human and identify its behavior, respectively. Blob vectors confirm that the extracted object from the frame of the environment is human or not, whereas the posture vector assures the behavior.

 This paper has been composed of four major sections. The following section summarizes the existing work and describing the core techniques used for human posture extraction. The third section discusses the proposed method, and consequently, in the next section, experimental work has been discussed. Afterward, the research work has been concluded with actual observations.

LITERATURE SURVEY

It has been observed that most of the uncertain situations that include human are majorly terrorist attacks. The primary step is to improve human behavior recognition by estimating human postures. Therefore, the preventative action to such attacks is to improve the security system through enhancing human posture apprehension (HPA). Currently, multiple analysts are employed to identify the suspect from the footage after attacks. So, in this accordance, a robust model is required to be developed. Therefore, various prior methodologies have been reviewed to analyze and overcome through posture apprehension problem.

According to I. Mikic et al., 2003, a built-in framework was modeled using ellipsoids and cylinders to automatically acquire a human body model from video streams with nonredundant parameters. Frames from the video were segmented. From the foreground silhouette, a human's body shape has reconstructed, acquiring the model in tracking algorithms. For this process, the Bayesian network was imposed for filtration, and the Kalman filter is employed to estimate the measurement through voxel data for tracking. In this way, action from different body sizes of dancing movements has been evaluated.

According to Voit M., Nickel K., and Stiefelhagen R., 2005, May, when the research moves around human-computer interaction, extraction of information is required for the human head. Therefore, in this methodology, the data is extracted from the head pose in a smart room. For different head poses, the horizontal head rotation has been performed by the people under the environment. The environment had been equipped with multiple cameras within the room to capture each person's facial view. The samples of rotated head poses have been collected to form a dataset to train the neural network for classification. If there is an estimation of head rotation, it has been observed

that more than one hypothesis of different estimations combined into one. This process results in a 50% mean error reduction for unknown head samples from different views.

Further, Yuan C. and Niemann H., 2001, presented the technique for pose estimation and recognition of three-dimensional objects from a single two-dimensional perspective by using the neural network-based system. It was an appearance-based neural approach. In this technique, an object has been represented through a principal component network that derives a feature vector. The NN classifier has identified the generated feature vector as an object using the resilient backpropagation algorithm. It generates four NN estimators from feature vectors to obtain pose parameters.

Further, A Rasekh, C. Chen, and Y. Lu, 2014, identified and recognized human activity in the world of medical sciences. Human activities were recognized using a 3D Smartphone accelerator (sensor) which collects time-series signals. Using these signals, 31 features have been generated in both the time and frequency domains. Passive learning methods, i.e., quadratic classifier, k-nearest neighbor algorithm, support vector machine, and artificial neural networks, are used to perform classification of the activities. In this technique, feature extraction and subset selection were performed for dimensionality reduction. Active learning algorithms are also used to reduce data labeling expenses.

Table 1. Related Literature Survey Summary.

First author	Year	Topic
C. Yuan	2001	Neural networks for the recognition and pose estimation of 3D objects from a single 2D perspective view
IVANA MIKIC	2003	Human Body Model Acquisition and Tracking using Voxel Data
Michael Voit	2007	Multi-view Head Pose Estimation using Neural Networks
Amin Rasekh	2014	Human Activity Recognition using Smartphone.
N. Zerrouki	2014	Automatic Classification of Human Body Postures Based on the Truncated SVD

The technique proposed by Zerrouki N. and Houacine A., 2014, automatically classify human body postures using SVD coefficient features. An input image is extracted from a fixed camera to classify through ANN, whereas background subtraction is performed for human body segmentation. A truncated SVD was applied to select significant magnitude coefficients. It includes a height-weight ratio of the human body in the set of features.

Table 2. Comparison of all the prior arts.

S. No.	Author's Name	Approach	Data Set	Limitation
1	IVAN MIKIC	Kalman filter and Bayesian Network on 3D Voxel data.	30 voxels data used.	Additional preprocessing steps are required to compute and reconstruct 3D voxel data from image data.
2	C. Yuan	PCA for feature extraction.	8 subjects were used to create the dataset.	Methodology fails to work in a heterogeneous environment.
3	Michael Voit	Neural network for estimating head poses.	Limited facial views used.	Require a smart room with multiple cameras for monitoring the movement of the head.
4	Amin Rasekh	3D Smartphone-based accelerometer for feature generation.	1393 samples of 3 persons were used.	Not simulated in a real environment.
5	N. Zerrouki	SVD (Singular Value Decomposition) for classification of postures.	A dataset of 4000 images.	Invariant features. Background subtraction method.

Table 2 shows the comparison of approaches and the datasets used in prior arts for HPA. Every researcher has their own technique to perform recognition and apprehension of postures. However, some significant limitations have been listed in the last column of the table. The following subsection gives a glimpse of the core terminologies found in posture apprehension research.

CORE TERMINOLOGIES

In this section, various core terminologies related to human posture have been identified. Core terminologies are the keywords in the field of research seen over the years. The different feature identification process used in HPA has been listed below in detail.

VOXEL DATA: Basically, the term voxel is used for the smallest unit in a 3D input similar to a pixel in a 2D input. A pixel is defined by 2 values in x and y, whereas a voxel is defined by 3 values as x, y, and z for a point. The concept of voxel data is to deploy reconstruction of voxel on input image frames/videos to extract the silhouette of a human (Mikic I., Trivedi M., Hunter E., and Cosman P., 2001). This will result in extra processing for the creation of voxel and the set of ellipsoids used as these shapes are closer to human body interpretation.

Principle Component Analysis (PCA): In HPA, the primary step is to extract the features. According to C.Yuan, PCA (Wold, S., Esbensen, K. and Geladi, P., 1987) can be applied to extract features. This method emerged to be very effective for extraction of the feature with the eigenvector technique. This derived feature vector by PCA is used on the NN classifier for the appearance-based classification of objects.

SVD: The acronym stands for Single Value Decomposition. These coefficients are generated as features of the object to perform automated classification. A fixed camera is deployed in the proposed structure (Golub G.H. and Reinsch C., 1971). After extracting images from the camera, a basic background removal technique is applied to segment the human from the image. The SVD is formulated by the magnitude of the coefficients, and the classification is performed using ANN.

ACCELEROMETER. It emerges as a device to an implant for various applications; one of them is the apprehension of human posture. Basically, the accelerometer is deployed in a smartphone; however, most smartphones have an accelerometer, and time series can be noted down. In the work proposed by Ravi N., Dandekar N., Mysore P., and Littman, M.L., 2005, time and frequency domain signals are captured to generate a feature vector for classification. K-nearest neighbor algorithm is used for the classification of human postures.

NEURAL NETWORK. To recognize human posture, estimation of a human's head has a significant role. The neural network is trained for various head poses. Whenever a head turns, the captured data from the camera is extracted from the camera and classified by the NN for a destined head pose results (Murphy-Chutorian E. and Trivedi M.M., 2008). The neural network is a feasible solution, but relying on multiple cameras is cumbersome. So, estimation of their head is absolutely not a feasible solution for HPA.

The core terminologies give a glimpse of processes and methods used in human behavior apprehension so far. These are considered as the basis for the development of the algorithm and presented in the next section.

MATERIALS AND METHODS

In the study, it has been observed that people can categorize people through postures in public places. Therefore, the postures are classified as suspicious or nonsuspicious because a person has a defined set of actions like seating and standing in public places like platforms, bus stands, shopping malls, technical campuses, organizations, and so on. The proposed methodology is to identify suspicious activities through a human posture in public places. The section will present the extraction of the featured posture of a human from the surrounding and classify them in suspicious and nonsuspicious behaviors.
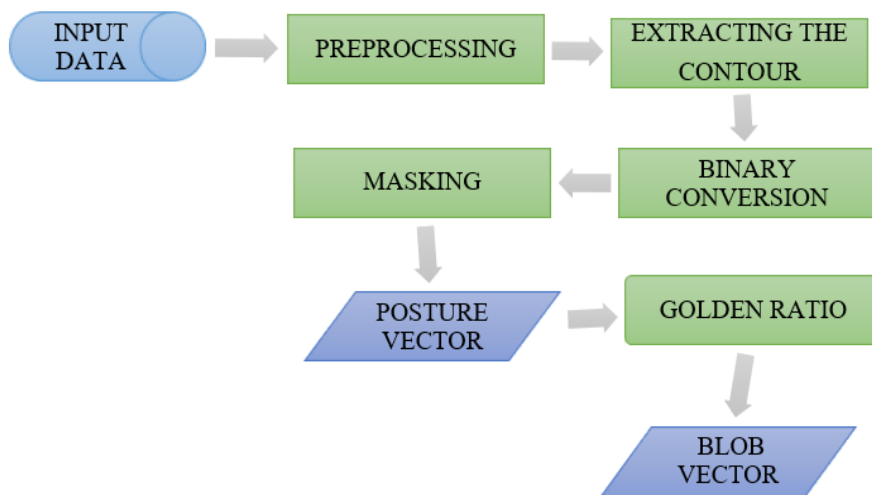

Figure 1. Flow chart of feature detection of HPA.

The proposed methodology is a combined approach of various processes to identify a human posture shown in Fig. 1. It shows the initial preprocessing steps required to preserve the image in the spatial domain by removing noise from the data frame. The following process after preprocessing is contours extraction. Contour represents the outline boundary of an object. The extracted contour is converted into binary to identify attributes through masks. Afterward, on the processed image in binary, golden ratio has been applied to segment the image to generate blob and posture vectors. All the processes are discussed in detail in later sections.   Golden ratio is applied on the filtered contour to identify a human blob for its authenticity and generating a human posture vector (Monika et al. 2015).

Proposed Methodology

In this section, the feature extraction procedure from on input sample is discussed in detail. As extraction of features is not a one-step task, all the phases are described in detail. Basically, 'EXTRACTION of FEATURES' refers to the technique of extracting featured attributes of the subject (here human) from the surroundings. It includes various significant tasks for feature extraction: preprocessing input, extracting the contour, binary conversion (applying edge detection), and finding out featured attributes (applying mask). Figure 1 above shows the complete method of extraction of a featured data from an input data set. The in-between steps involved are preprocessing input to filter noise and distortion from input data. The next phase in this process is the extraction of contour, that is, extracting the desired object silhouette. In the next phase, edges are filtered out. Afterward, masks are applied to extract the desired feature values. Finally, the featured data has been filtered to save minimum featured attributes and generate the blob vector to identify that the input set is a human.

Preprocessing of an input sample is applied to get a filtered input sample to efficiently process the extracted features. For effective preprocessing of input, the image is sampled to dimensionality reduction and filtering out noise and distortion from the input. suppose the input is I with dimension as x and y, which are values in x and y directions respectively, and R, G, and B are three planes of the input. In this process, the RGB image is sampled on a threshold for every plane, represented by the equation below as equation 1.

$$P[x, y] = I[x, y] - [\theta1 * R + \theta2 * G + \theta3 * B] \tag{1}$$

Here, in Equation 1, $P[x, y]$ is processed data, and $\theta1, \theta2, \theta3$ are the threshold for R, G, and B planes. The preprocessing results of the seating and standing dataset are shown in Table 4 in Column 2.

EXTRACTION of CONTOUR is the following process in feature extraction to distinguish between the subject (human) and background. The purpose of this phase is to differentiate between background and foreground objects. A threshold value at 128 is fixed to distinguish between a background and a human to optimize the dataset. The equation to extract the contour from the preprocesses sample is shown in equation 2.

$$H[x, y] = 1 \; if \; P[x, y] = \theta \; otherwise \; 0 \tag{2}$$

Here, in the equation, $H[x, y]$ is the extracted contour from preprocessed input $P[x, y]$, and θ is the threshold. In this equation, if the processed input has a value similar to the threshold, we get those values. Otherwise, the values are set to zero, and this process will result in $H[x, y]$ where subjects features exist, and background is null. The contour extraction results of the seating and standing dataset are shown in table 4 in column 3.

BINAY CONVERSION: The next phase is reducing the dimension by applying edge detection to minimize and get optimal features from the contour. A pair of convolution mask is used, as shown as follows:

$$bx = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix}, by = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} \tag{3}$$

with a gradient strength of $G = \sqrt{gx^2 + gy^2}$ $\qquad\qquad$ (4)

It is applied over the contour $H[x, y]$ to generate a binary conversion of the extracted contour. Further, various masks have been used to extract the features from the edge detection result produced by binary conversion; all masks are shown in Figure 2.

MASKING: It is applied to extract the featured attribute from the binary image. In the proposed approach, six different masks of 9×9 to extract the featured attributes from the binary image generated in the previous phase have been applied. The masks basically extract a left arc, a right arc, and a horizontal and vertical continuous edge to get featuresfrom a human sample, including its complete body. The masks are shown in Fig.2; Fig.2(a) and (c) show the mask that finds out left arcs, whereas Fig.2(b) and (d) extract right arcs, and lastly, Fig. 2(e) and (f) masks are applied to extract horizontal and vertical edges. The masking results on seating and standing postures are shown in Table 4 in the fourth column.

(a)　　　　　　(b)　　　　　　(c)

(d)　　　　　　(e)　　　　　　(d)
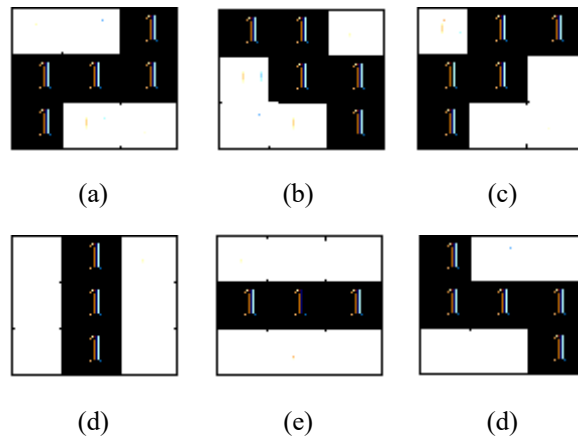
Figure 2. 6 Different masks are used to extract features which are shown above from (a) to (f).

POSTURE VECTOR: The proposed system is to identify the postures for identifying suspicious behaviors. Therefore, to train the NN, a "Posture Vector" is created where each input sample is filtered from the above phases. A feature vector of "2408" elements from each training sample to train the network for human postures ranges between 0 and 1 is generated. It has been shown through the graph in column 2 of Table 5.

BLOB VECTOR: The last step is to identify blobs from the filtered sample of masking to generate the input vector set for training the blob samples to identify a human. The extraction of the blob from the masking results is applied by segmenting the sample binary masked input in the defined golden ratio proportion. The blob (head) vector graph for standing and seating postures is shown in column third of Table 5.

In the next section, the dataset used for the experimental work has been discussed with results in detail.

EXPERIMENTAL RESULTS & DISCUSSION

In this section, the datasets used for the experimental work have been discussed in detail. The preprocessing results of the proposed methodology have been demonstrated. This section has been further divided into three subsections a dataset, processing results, and analysis.

DATASET: Dataset is a collection of samples for training, testing, and validation of a model. It consists of several samples, such as containing 'n' number of features to perform efficient classification, clustering, and regression to fulfil various objectives. The details on the dataset used in this paper can be verified from Mehta, Monika 2015. and Monika Mehta and Madhulika Bhadauria, 2021. As classification is the proposed system's primary task, the dataset needs to be designed carefully. To create a training dataset, a specified environment has been set up. Other conditions are also considered to generate a high-quality input sample. Therefore, a 10×10 feet room has been selected. A medium-light source is projected to control brightness and contrast on the subject (human). The subject is supposed to wear a costume while posing for different postures for the purpose of uniqueness.

Table 3. Dataset specification of the proposed system.

Type of dataset	Pose with NO Actions	Pose with Differential Action
Standing Human Pose Training	20	17
Seating Human Pose Training	14	10

For more closeness to real-time data, the input training samples are added with some gestures pose like a person holding a hand grill in metros. For seating posture, a Davenport is used to get seating postures of a human similar to a human seating in the bus, in metros, or on platforms. Also, it is observed that, sometimes, people are seated on the ground, so we had included some sample training dataset related to this posture. We had made 50+ samples of the dataset for training in a defined set of the environment using a high definition DSLR camera of 24.2 Megapixels. We had created around 40 samples of humans standing posture with some actions to train the network and 25 samples of a seating pose, including differential actions to train the network as listed in Table 3. Figure 3 shows the sample dataset of seating and standing postures during the initial phase of processing.

(a)   (b)   (c)   (d)
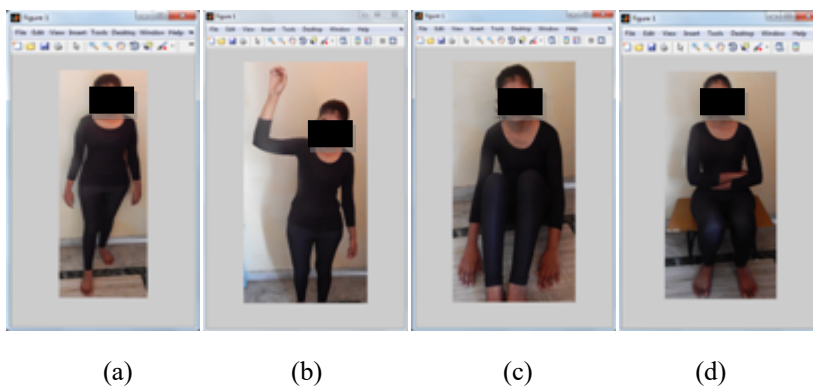
Figure 3. The samples of the training dataset.

PREPROCESSING RESULTS

In this section, the in-between processing results of the proposed methodology, such as preprocessing of input, extraction of contour, and masking results on different samples, have been demonstrated in Table 4. Further, in Table 5, the input histogram of the postures used in Table 4 has been demonstrated with its posture vector and blob vector containing posture and blob feature points, respectively. The histogram shows that the input spectrum ranges from 0 to 255, whereas the posture vector and blob vector contain 2408 elements ranges from 0 to 1.

Table 4. Results of feature extraction phases.
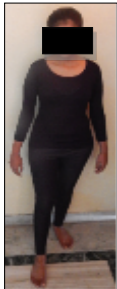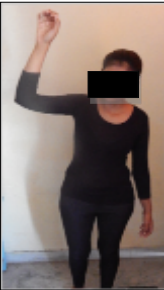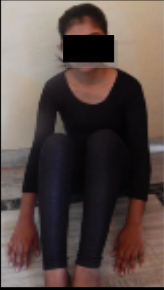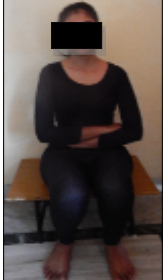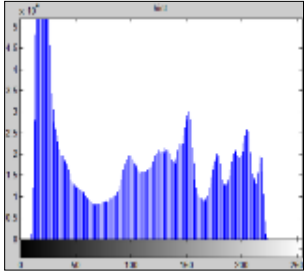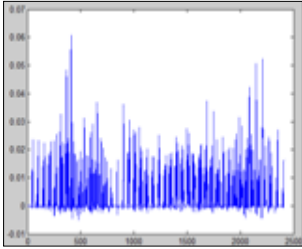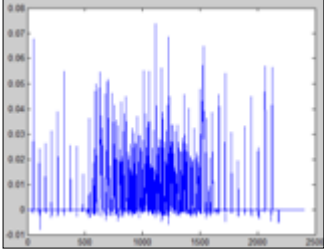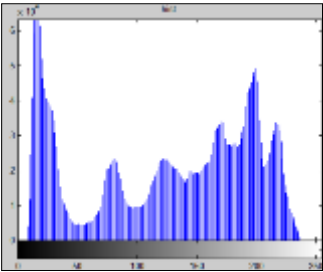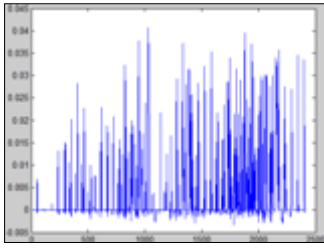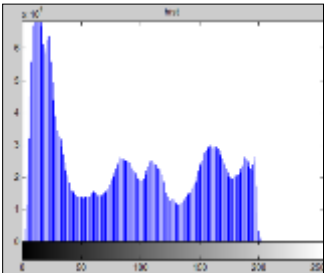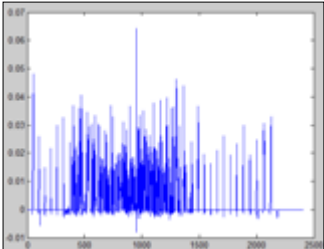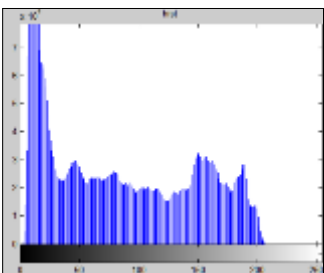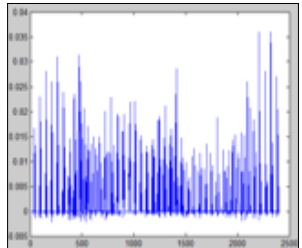
	PHASE	INPUT	PRE- PROCESSING RESULTS	EXTRACTION of CONTOUR	MASKING
STANDING POSTURES	SAMPLE 1				
	SAMPLE 2				
SEATING POSTURES	SAMPLE 1				
	SAMPLE 2				

Table 5. Input Histogram and Training vector graph.

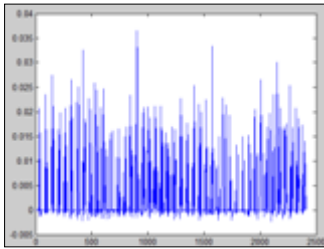|	PHASE	HISTO-GRAM	POSTURE VECTOR	BLOB VECTOR
S T A N D I N G P O S T U R E S	**S A M P L E 1**			
	S A M P L E 2			
S E A T I N G P O S T U R E S	**S A M P L E 1**			
	S A M P L E 2			

RESULT ANALYSIS

Extraction of features from the input sample is a time-consuming process; however, a reasonable rate of extraction resulting in effective learning in network training has been attained. All the results related to the extraction of features are presented in Table 4 through few samples. The sample analysis is identified by calculating the range minimum and maximum values given in Equations 5 and 6. The terms P_V and B_V in the equation are for posture vector and blob vector, respectively.

$$minimum\_posture\_value = min(\textstyle\sum_i^{2408} P\_V(i)) \tag{5 (i)}$$

$$minimum\_blob\_value = min(\textstyle\sum_i^{2408} B\_V(i)) \tag{5 (ii)}$$

$$maximum\_posture\_value = max(\textstyle\sum_i^{2408} P\_V(i)) \tag{6 (i)}$$

$$maximum\_blob\_value = max(\textstyle\sum_i^{2408} B\_V(i)) \tag{6 (ii)}$$

Basically, the analysis result of the dataset shows that the feature vector of samples is used to train, ranging from 0 to 1, which signifies that it is elementary to store and manipulate. If we look at Table 5, the standing posture sample 1 and sample 2 are different, but the value range is almost near for blob and posture. Similarly, seating postures of Human positions are different, for example, in one sample, subject seating on the ground, while, in another sample, subject being on an object, but the samples match nearby values. Table 6 shows that the final feature vector generated through the proposed methodology has been operationally feasible. This feature vector is easy to store, manipulate, and operate, and the values of the same type of postures are almost similar.

Table 6. Analysis of samples.

Type of dataset	Blob Value		Posture Value	
	Minimum Value	Maximum Value	Minimum Value	Maximum Value
Standing Human Pose Sample 1	-0.0040	0.0513	-0.0034	0.0406
Standing Human Pose Sample 2	-0.0039	0.0637	-0.0033	0.0457
Seating Human Pose Sample 1	-0.0080	0.0643	-0.0023	0.0365
Seating Human Pose Sample 2	-0.0110	0.0815	-0.0022	0.0336

CONCLUSION

Presently, public place security is performed extensively using metal detectors, x-ray scanners, and CCTV cameras. Public place security is mainly dependent on video recorded through CCTV cameras. Human posture has been identified as the most potent method for human behavior recognition in public places from videos. The paper presents a novel approach that generates featured attribute sets of a human as posture and blob vector. These attribute vectors are used for human authentication and posture identification. These vectors are correlated and easy to use with machine learning-based models for classification. The lower numerical feature set needs less computation in training and testing of the model. In the future, the authors will test this method with Machine Learning Model on the video-based dataset.

REFERENCES

A. Rasekh, C. Chen, and Y. Lu, 2014. Human activity recognition using Smartphone. CoRR, vol. abs/1401.8212.

Chen, Liang-Hua, Li-Yun Wang, and Chih-Wen Su 2014. Human detection in surveillance video. International Journal of Pattern Recognition and Artificial Intelligence 28.02:1455003.

Golub, G.H. and Reinsch, C., 1971. Singular value decomposition and least squares solutions. In Linear Algebra (134-151). Springer, Berlin, Heidelberg.

I. Mikic et al. 2003. Human body Model Acquisition and Tracking Using Voxel Data. International Journal of Computer Science Vision. 53(3): 199-223.

Kadamvari et al.2015, A Comparative Study of Pedestrian Identification Approaches for Traffic Light Simulation. International Journal of Computer Science and Technology. 5(4):69-72.

Mehta, Monika 2015. Human Posture Apprehension, Lap Publishing, 978-3659767258.

Mikic, I., Trivedi, M., Hunter, E. and Cosman, P., 2001, December. Articulated body posture estimation from multi-camera voxel data. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001 (Vol. 1, pp. I-I). IEEE.

Monika and Madhulika2014. A perlustration of Human Apprehension and Behavior Accedence in surroundings. Journal of Multidisciplinary Engineering & Technology. 8(1):38-46.

Monika et al., 2015. Detecting Human from Environment: Using AI for embellishing preprocessing step of Human Accedence. International Journal of Computer Science and Technology (IJCSET). 5(4):76-80.

Monika Mehta * and Madhulika Bhadauria, 2021, "Performance Evaluation of Neural Network for Human Classification using Blob Dataset", Recent Advances in Computer Science and Communications 2021(14), p.1594-1604.

Murphy-Chutorian, E. and Trivedi, M.M., 2008. Head pose estimation in computer vision: A survey. IEEE transactions on pattern analysis and machine intelligence, 31(4):607-626.

Official website of Delhi Metro Rail Corporation. [Online]. Available:www.delhimetrorail.com/ Accessed: 15-May-2020**.**

Ravi, N., Dandekar, N., Mysore, P. and Littman, M.L., 2005, July. Activity recognition from accelerometer data. In AAAI. 5(2005):1541-1546.

Voit, M., Nickel, K. and Stiefelhagen, R., 2005, May. Multi-view head pose estimation using neural networks. In The 2nd Canadian Conference on Computer and Robot Vision (CRV'05) (pp. 347-352). IEEE.

Wang, Z. and Nagai, Y., 2018. Research into the Improvement of Museum Visitor's Emotional Response Levels to Artificial Lighting Designs Based on Interdisciplinary Creativity. on Creativity and Innovation 2018, p.330.

Wold, S., Esbensen, K. and Geladi, P., 1987. Principal component analysis. Chemometrics and intelligent laboratory systems. 2(1-3):37-52.

Yuan, C. and Niemann, H., 2001. Neural networks for the recognition and pose estimation of 3D objects from a single 2D perspective view. Image and Vision Computing. 19(9-10):585-592.

Zerrouki, N. and Houacine, A., 2014. Automatic classification of human body postures based on the truncated SVD. Journal of Advances in Computer Networks. 2(1):58-62.