Automatic Human Emotion Recognition in Surveillance Video

J. Arunnehru and M. Kalaiselvi Geetha

Abstract Recognition and study of human emotions have fascinated a lot of attention in the past two decades and have been researched broadly in the field of computer vision. The recognition of complete-body expressions is significantly harder, because the pattern of the human pose has additional degrees of self-determination than the face alone, and its overall shape varies robustly during articulated motion. This chapter presents a method for emotion recognition based on the gesture dynamics features extracted from the foreground object to represent various levels of a person's posture. The experiments are carried out using publicly available emotion recognition dataset, and the extracted motion feature set is modeled by support vector machines (SVM), Naïve Bayes, and dynamic time wrapping (DTW) which are used to classify the human emotions. Experimental results show that DTW is efficient in recognizing the human emotion with an overall recognition accuracy of 93.39 %, when compared to SVM and Naïve Bayes.

Keywords Emotion recognition · Dynamic time wrapping · Support vector machines · Naïve Bayes · Object detection

1 Introduction

Human–Computer interaction is mounting its attention nowadays. In order to put some prominence on socializing computer with human, understanding the human body gestures and visual cues of an individual is a need [1]. It allows a system to understand the expressions of humans in turn, enhancing its effectiveness in performing various tasks. It serves as a measurement system for behavioral science,

J. Arunnehru (✉) · M. Kalaiselvi Geetha
Department of CSE, Annamalai University, Chidambaram, India
e-mail: arunnehru.aucse@gmail.com

M. Kalaiselvi Geetha
e-mail: geesiv@gmail.com

N. Dey and V. Santhi (eds.), *Intelligent Techniques in Signal Processing for Multimedia Security*, Studies in Computational Intelligence 660,
DOI 10.1007/978-3-319-44790-2_15

and socially intelligent software tools can be accomplished. Recognition of gesture-based human emotions has attracted a lot of attention in the precedent two decades and has been researched broadly in the field of computer vision. The recognition of complete-body expressions is significantly difficult to express, because the semantic pattern of the human pose has additional degrees of range than the face alone, and its overall shape varies robustly during articulated motion [2, 3].

Human emotion recognition is having a wide range of human–computer interaction (HCI)-related applications. Future shopping is going to be more analytical in the near future using video to track emotions of the buyers. Future retail will avoid silly queries from the customers and combine business, technology, human behavior, and psychology using HCI. For example, if a customer looks frustrated in interactive monitors, then the retailer could understand that something needs to be done for him. Thus, HCI enables the marketers to analyze the actual customer behavior in emotions in real time than to process-biased answers or survey questions.

This chapter investigates recognizing emotions from gestures which is a challenging task. Automatic human emotion recognition system could offer the marketers the ability to interact with the customers in real time, facilitate better decisions, and to be effective in providing service to customers.

Human emotions emerge differently on the same external stimulus through individual standards in the individual style [4]. According to gender, age, society, or residential areas, the emotion expression can be different on the same multimedia information. The most important challenge in HCI-based system is to provide the capability to computers to evaluate human emotions strongly. The immense potential applications of human emotion recognition, such as interactive learning systems, consumer care, web cinema, safety and video surveillance, just to name a few, provide increased need of a strong human emotion recognition. A massive amount of research has been done on human emotions, and it is almost impossible to identify all of it. One can categorize the proposed methods based on features extraction. Action recognition systems can be divided into three categories: (1) walking, (2) sitting, and (3) jumping.

Generally, human emotions consist of 3 principal emotions: happy, angry, and fearful. Remaining emotions are considered to be combinations of these primary emotions. The outcome reported the efficacy of combining the visual information into single framework. The majority of the multimodal systems better the modal approaches for emotion recognition applications. To understand the emotions from human pose motion in normal environments can be extremely hard as body movements are almost unconstrained. This makes it complicated to train emotion recognizer's which are strong as much as necessary to endure this kind of real-world inconsistency problems.

Further, an emotion recognition approach must be able to discriminate between emotions like happy, angry and fearful, combined with activities like walking, sitting, and jumping performed by an individual person. This is not an easy task, since certain movements such as going from happy-walking posture to angry-walking posture have strong similarities. Hence, the problem of human emotion recognition is viewed as motion gesture in this proposed work. This chapter analyzes the emotion recognition problem as vision-based gesture recognition. In vision-based gesture recognition, the procedure is carried at four steps, viz. human detection, human tracking, emotion recognition, and then a complex activity assessment to evaluate happy, angry, and fearful emotions.

Automatic emotion recognition has turned into a significant research area in computer vision for last few decades with applications in video surveillance, sport event analysis, human–computer interaction, computer-aided games, etc. Recent studies on visual analysis of human body pose movement reveal that the human activities vary from other motion movements. Speech emotion recognition is achieved through untainted sound processing without linguistic information. Anagnostopoulos et al. [5] proposed speech emotion recognition through processing approaches that include the separation of the speech signal, and speech features are extracted for the emotion classification. For capturing the information about emotion from audio and facial features, auto associative neural network (AANN) models [6] are considered in video for emotion recognition. Although several automatic action recognition systems have considered the use of both gesture and facial expressions, relatively few efforts have focused on emotion recognition using both modalities. The emotion recognition is based on the combination of facial expressions and speech data [7] and facial expressions and gesture [8]. Karpouzis et al. [9] presented a multi-cue framework based on facial, oral, and bodily expressions to model emotional states, but the synthesis of modalities is modeled at the point of facial expressions and speech information only. The effort–shape analysis [10] was used to illustrate the movement style distinctiveness connected with each of the objective emotions on knocking movement, and the target emotions are angry, anxious, content, joyful, proud, and sad. Castellano et al. [1] proposed human emotion recognition of four performed emotional states (angry, joy, fearful, and sadness) based on the psychiatry of body pose movement and gesture expressivity. The various methods from psychology focus on the relationships between action and emotion behavior, investigating expressive body movements [11, 12]. The automatic analyses of emotions in multimedia records are helpful for indexing and retrieving the multimedia information based on emotion-specific information [13].

Kapoor et al. [14] presented a method based on correlation between body posture and aggravation in a computer-based training environment. Kapur et al.'s [15] four basic emotions may possibly be automatically distinguished from statistical measures of motion's dynamics. Balomenos et al. [16] proposed a technique by fusing the facial expressions and body gestures for the recognition of six classical emotions. Camurri et al. [17] presented a method based on human full-body movement to discriminate the expressive gestures. In particular, they recognized motion cues similar to time duration, contraction index, magnitude of motion, and motion confidence. From these motion cues, they defined an intelligent classifier that has the ability to discriminate four emotions, namely angry, fear, sorrow, and joy. Zhaojun Yang et al. [18] presented a graph-based approach to identify the gesture's dynamic pattern and emotion from body motion cues in common interpersonal interactions, where dynamic patterns for particular emotions from a weighted graph-based method improved separation among distinct emotion classes in order to maintain fewer inconsistencies within the similar emotion class. Zaboleeva-Zotova et al. [19] presented a new methodology for recognition of human emotions based on analysis of body distinguishing gestures and poses; the characteristics of body pose movements are modeled with linguistic variables for sequential activities.

From the literature, it is clear that emotions are mostly expressed as mental or psychological states [20], which are the most important human cognitive features that attract life by relations processed in the human race [21]. Mental states like happy, angry, jealousy, fear, envy, indignation, embarrassment are classified and referred to "Emotion." Recently, artificial intelligence researchers have measured the significance of adding emotion to computers, which prioritize the primary human motive and direct their processing for full expression of emotions [22]. However, one can easily see emotion from an image sequence than from a still image. One reason for that is the information from the image sequence contains appearance and motion feature. Due to this reason, the analysis of dynamic image sequences has become very attractive in computer vision, virtual reality [23], and cyborgs [24] field; this has motivated to analyze the gesture expressions from a dynamic image sequence in the proposed work.

2 Emotion Recognition Framework—Overview

For better understanding of the underlying study of this chapter, a real-life scenario for emotion recognition is experimented. Dataset comprising human emotion performed by several actors in static environment is used for experimental purpose. The proposed work employs ideal features, extraction approaches, and classifier which reveal promising outcomes. The overview of the proposed approach is shown in Fig. 1.

This chapter deals with human emotion recognition, which aims to discriminate different emotions based on actions from the video sequences. The proposed method is evaluated using emotion recognition dataset [25], considering emotions such as happy, angry, and fearful with activities like walking, sitting, and jumping. Background subtraction technique is applied to current frame in order to obtain the foreground object. Thus, the motion and shape features are computed and chosen as

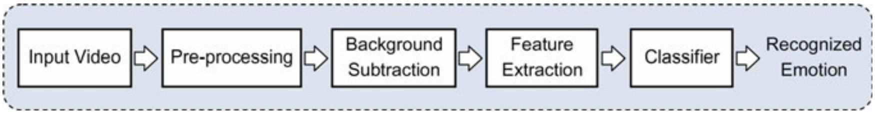


Fig. 1 Overview of the proposed emotion recognition approach

a feature set. The extracted feature set is fed to the SVM, Naïve Bayes, and DTW classifiers for classification.

2.1 Feature Description

The extraction of selective characteristic is the most essential crisis in human emotion and activity recognition which represents the momentous information that is necessary for further study. To identify a person's movement across image sequences, background subtraction approach is widely used. The foreground object is detected by subtracting the current frame from the reference frame. The foreground images are obtained by applying thresholds to reduce pixel modification due to camera noise and changes in illumination conditions. This process is really adaptive to perceive the motion region equivalent to moving objects in static scenes and better quality for extracting significant feature pixels.

The foreground object is obtained by simply subtracting the current frame at time $I(x,y,t)$ from background reference frame at time $B(x,y,t)$ on a pixel-by-pixel basis. The extracted foreground object $F(x,y,t)$ is considered as the region of interest (ROI) or minimum bounding box. Figure 2a and Fig. 2b illustrate the background frame and current frame of the happy-walk emotion dataset. The resulting foreground object is shown in Fig. 2c. The sample foreground objects for different human emotions are extracted from the emotion dataset [25] and are shown in Fig. 3. From the figure, it is clearly indicated that the foreground object (shape) represents the different gesture dynamics of action and emotion. The foreground object $F(x,y,t)$ is calculated using

$$\begin{aligned} &F(x,y,t) = |I(x,y,t) - B(x,y,t)| \\ &1 \leq x \leq w, 1 \leq y \leq h \end{aligned} \tag{1}$$

$$F(x,y,t) = \begin{cases} 1, & \text{if } I(x,y,t) < t \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where w—width of the foreground object, h—height of the foreground object, and threshold $t = 30$ is to be considered in this work.

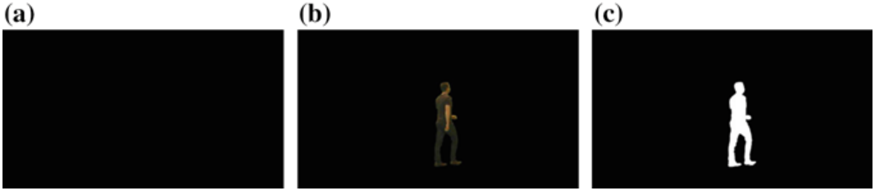


Fig. 2 **a** Background model, **b** current frame, **c** foreground object obtained from (**a**, **b**)

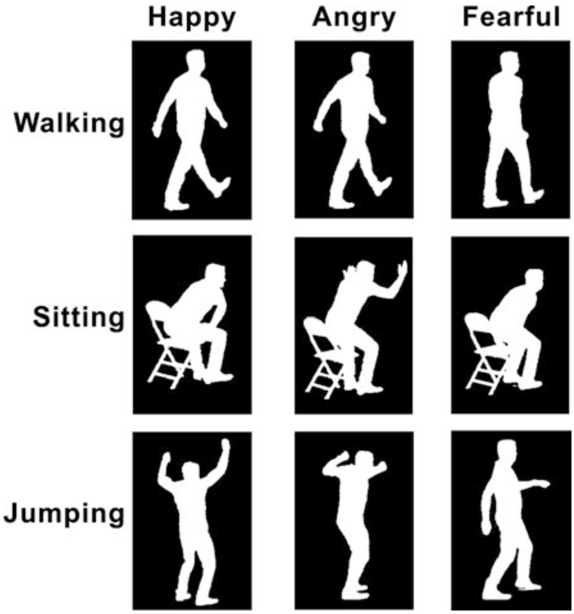


Fig. 3 Sample foreground objects from emotion recognition dataset

To recognize the human emotion, motion and shape information is an essential signal usually extracted from video sequences. Distance, speed, orientation, elongation, solidity, and rectangularity measures are compact representation of motion and shape information, since much valuable information is retained in this measure. The distance and speed are extracted from the successive frame that consists of foreground object motion information only. The orientation, elongation, solidity, and rectangularity are extracted from the foreground object that consists of shape data only.

Distance is a measure between successive frame objects by finding the centroid of the object.

$$\text{Centroid} = \left(\text{ROI}_{\text{width}/2}, \text{ROI}_{\text{height}/2}\right) \tag{3}$$

Euclidean distance measure is used to calculate the distance between (x_1, y_1) and (x_2, y_2) centroid points.

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{4}$$

where D—is the distance, (x_1, y_1)—centroid point on frame t, (x_2, y_2)—centroid point on frame $t + 1$.

Speed is an essential cue to determine the object's motion as fast or slow. Speed is defined as distance divided by time. Distance is directly proportional to velocity when time is constant. After finding the displacement of the object, speed is calculated using

$$S = \frac{D}{t} \tag{5}$$

where S—is the speed in m\s, D—is the distance travelled in pixels, t—is the time taken, $t = 0.08$ (25 fps\2—two frames used for distance measure).

Orientation is a measure of observant angle between the x-axis and the y-axis of the ellipse that has the similar second-moments as the region. The obtained cost is in degrees, ranging from −90° to 90°. Figure 4 illustrates an image area, and its corresponding ellipse and same ellipse with the solid red lines represents the axes; the orientation is the angle between the horizontal line axis and the vertical line axis.

Elongation (Elo) is also called as minimum bounding rectangle or minimum bounding box of an arbitrary shape. In arbitrary shape, eccentricity is the relation of the length L and width W of minimum bounding rectangle of the shape at some set of orientations as shown in Fig. 5 Elongation is a measure of values range from [0,1].

$$\text{Elo} = 1 - W/L \tag{6}$$

A regular shape in all axes such as a circle or square will have an elongation value of 0, whereas shapes with large aspect ratio will have an elongation closer to 1.

Solidity characterizes the extent corners to which the shape can be represented by convex or concave, and it is defined by

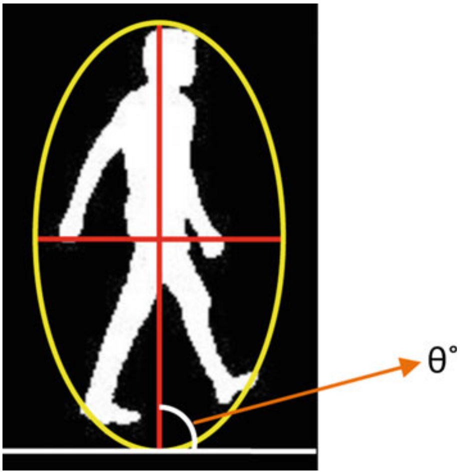


Fig. 4 Orientation of the angle between x-axis and y-axis moments

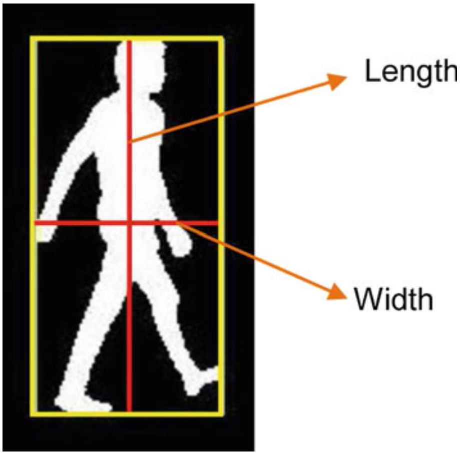


Fig. 5 Minimum bounding box and corresponding parameters for elongation

$$\text{Solidity} = A_{\text{S}}/H \tag{7}$$

where A_{s} is the shape region area and H is the shape area of convex hull. The solidity of a convex shape is always 1 as shown in Fig. 6.

Rectangularity describes the rectangular shape is how much it fills its minimum bounding rectangle as shown in Fig. 7

$$\text{Rectangularity} = A_{\text{s}}/A_{\text{R}} \tag{8}$$

where A_{S} is the shape area, A_{R} is the minimum bounding rectangle area.

The motion and shape cues such as distance, speed, orientation, elongation, solidity, and rectangularity are represented as six-dimensional feature vectors to depict the emotion and action. The extracted features are fed to the SVM, Naïve Bayes, and DTW for human emotion recognition.

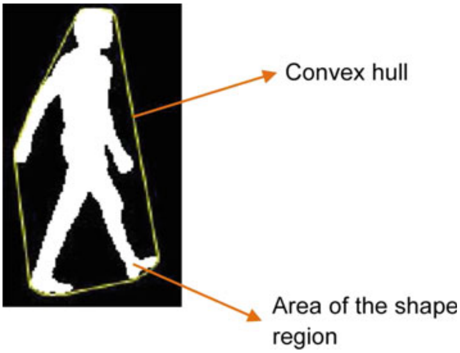


Fig. 6 Solidity of the shape is represented by shape area and convex hull area

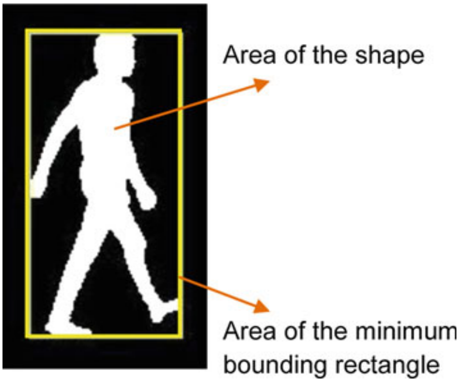


Fig. 7 Rectangularity of the shape is represented by shape area and minimum bounding rectangle area

3 Support Vector Machines (SVM)

Support vector machine (SVM) is a well popular technique for classification in visual pattern recognition [27]. The SVM is most widely used in kernel learning algorithm. It achieves reasonable vital pattern recognition performance in optimization theory. Classification tasks are typically involved with training and testing data. The training data are separated by $(x_1, y_1), (x_2, y_2), \ldots (x_m, y_m)$ into two classes, where $x_i \in \mathbb{R}^n$ contains n-dimensional feature vector, and $y_i \in \{+1, -1\}$ are the class labels. The aim of SVM is to generate a model which predicts the target value from testing set. In binary classification, the hyper plane $w.x + b = 0$, where $w \in \mathbb{R}^{\mathrm{n}}$, $b \in \mathbb{R}^{\mathrm{n}}$ is used to separate the two classes in some space $\mathbb{Z}$ [28]. The maximum margin is given by $M = 2/||w||$ as shown in Fig. 8. The minimization problem is solved by using Lagrange multipliers $\alpha_i (i = 1, \ldots m)$ where w and b are optimal values obtained from Eq. 9.

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b\right) \quad (9)$$

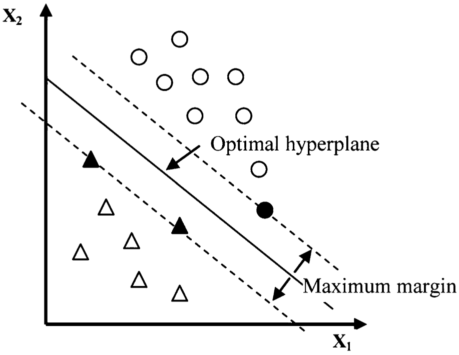


Fig. 8 Illustration of hyperplane in linear SVM

Table 1 Types of SVM inner product kernels

Types of kernels	Inner product kernel $K(x^T, x_i)$	Details
Polynomial	$(x^T x_i + 1)^p$	Where x is input patterns x_i is support vectors σ^2 is variance $1 \leq i \leq N_s$ N_s is number of support vectors β_0, β_1 are constant values p is degree of the polynomial
Gaussian	$\exp\left[-\frac{\|x^T - x_i\|^2}{2\sigma^2}\right]$	
Sigmoid	$\tanh(\beta_0(x^T x_i) + \beta_1$	

The non-negative slack variables ξ_i are used to maximize margin and minimize the training error. The soft margin classifier is obtained by optimizing by Eqs. 10 and 11.

$$\min_{\omega,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \tag{10}$$

$$y_i\left(w^T \phi(x_i) + b \geq 1 - \xi_i, \xi_i \geq 0\right. \tag{11}$$

If the training data are not linearly separable, the input space mapped into high-dimensional space with kernel function $K(x_i, x_j) = \phi(x_i).\phi(x_j)$ is explained in [29]. There are several SVM kernel functions as given in Table 1.

4 Naive Bayes

In recent years, researchers in pattern recognition, machine learning, and classification have been concerned with naive Bayesian classifiers. The naive Bayes algorithm makes use of Bayes theorem, which is a formula that determines probability by estimating the frequency of values and mixture of values. A naive Bayes is a simple probabilistic-based classifier, which can able to predict the probabilities of the membership class [28]. The naive Bayesian classifier is simple and computationally efficient learning algorithm with theoretical roots in the Bayes theorem. The Bayes theorem states:

- Let $A_1, A_2, \ldots, A_L$ be mutually exclusive events whose union has probability one. That is, $\sum_{i=1}^{L} P(A_i) = 1$
- Let the probabilities $P(A_i)$ be known.
- Let B be an event for which the conditional probability of B given A_i, $P(B|A_i)$ is known for each A_i.

Then:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{L} P(B|A_i)P(A_i)} \tag{12}$$

The probabilities $P(A_i|B)$ reflect updated or revised beliefs about A_i, in light of the knowledge that B has occurred. Once the probabilities have been estimated, the class is predicted by identifying the most probable one.

5 Dynamic Time Wrapping (DTW)

Dynamic time warping (DTW) is an essential method for measuring resemblance between two temporal sequences which may vary in time [26]. For example, similarities in consecutive patterns might be identified using DTW, even if single person was running faster than the other one, or if there were accelerations and decelerations through the path of a scrutiny. DTW has been efficient to temporal sequences of video, which can be turned into a linear sequence and analyzed with DTW. Figure 9 shows the time association of two time-dependent sequences.

The theory of DTW is to measure two (time-dependent) sequences $x := (x_1, x_2, \ldots, x_n)$ of length $N \in \mathrm{N}$ and $Y := (y_1, y_2 \ldots, y_m)$ of length $M \in N$. These sequences possibly will be distinct signals (time series), feature sequences sampled at equidistant points in time, and the feature space is denoted by F. Then $x_n, y_m \in F$ for $n \in [1:N]$ and $m \in [1:m]$. To evaluate two diverse feature vectors $x, y \in F$, one desires a partial cost measure, sometimes also referred to as local distance measure, which is defined as a function

$$C : F \times F \rightarrow R \geq 0 \tag{13}$$

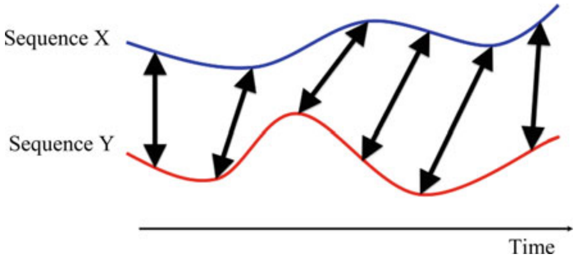


Fig. 9 Time alignment for two time-dependent sequences

6 Emotion Dataset

Emotion dataset (University of York) is a publicly available dataset [25], containing four different emotions (happy, angry, fear, and sad) performed by 25 actors. The sequences were taken over static (black) background with the frame size of 1920 × 1080 pixels at a rate of 25 fps. For each emotion, actors are performed five different actions: walking, jumping, box picking, box dropping, and sitting, having an approximate length of 15 s of video. In this work, three emotions (happy, angry, and fear) and three actions (walking, jumping, and sitting) of 10 persons (male and female) are used for experimental purpose. Table 2 shows the properties of emotion recognition dataset.

In Figure 10, first row shows the walking action recognition with three different emotions, happy, angry, and fearful. Second row shows the sitting action recognition with three different emotions, happy, angry, and fearful, and the third row shows jumping with three different emotions, happy, angry, and fearful. For each approximate length of 15 s of video obtained, 90 data records are considered for experimental purpose. In this work, 10 persons are taken randomly from emotion dataset for evaluation. The samples are divided into training set of 5 persons and testing set of 5 persons.

7 Performance Measure

Table 3 illustrates a confusion matrix for a human emotion recognition problem having true positive, false positive, true negative, and false negative class values. If the classifier predicts correct response of class at each instance, it is counted as "success," if not, it is an "error." The overall performance of the classifier is obtained by error rate, which is a proportion of the errors made over the whole set

Table 2 Properties of 10 subjects in emotion recognition dataset

		Happy (s)			Angry (s)			Fearful (s)		
Subject	Gender	Walk	Sit	Jump	Walk	Sit	Jump	Walk	Sit	Jump
1	Male	3	5	7	2	5	6	5	2	4
2	Female	3	4	4	2	3	2	5	4	3
3	Male	3	7	6	2	8	4	2	6	7
4	Female	4	3	3	4	3	4	3	3	6
5	Male	4	6	4	5	8	4	5	7	3
6	Female	3	5	5	3	4	4	6	5	3
7	Female	4	7	4	3	2	3	2	6	8
8	Male	3	8	4	2	3	4	2	2	4
9	Female	3	5	3	3	4	3	2	6	5
10	Male	3	6	4	3	3	4	2	6	5

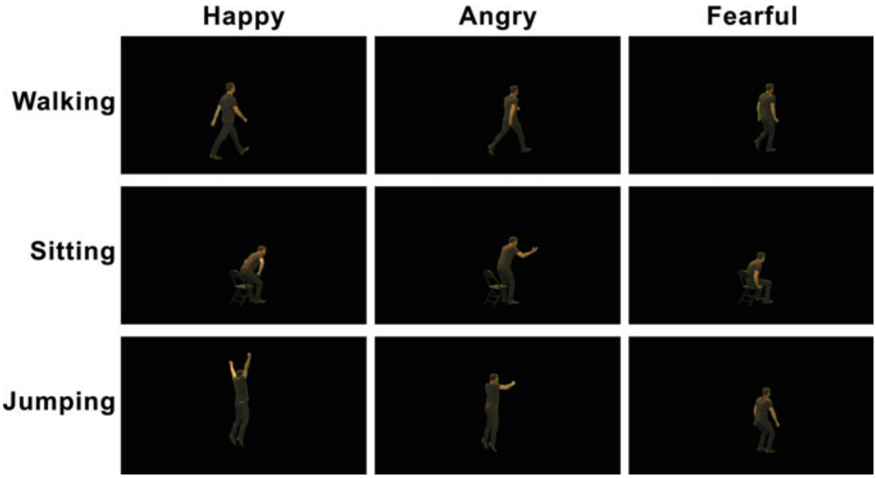


Fig. 10 Sample frames from the emotion dataset

Table 3 Confusion matrix

Actual	Predicted	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

of instances. From the confusion matrix, it is possible to extract a statistical metrics (recall, precision, F-measure, and accuracy) for measuring the performance of classification systems and is defined as follows:

Recall (R) or sensitivity is a ratio between correctly labeled instances and total instances in the class. It has an ability to measure the prediction model and is also called as true positive rate. It is defined by:

$$\text{Recall (R)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

where TP and FN are the numbers of true positive and false negative predictions for the particular class. TP + FN is the total number of test examples of the particular class.

Precision (P) or detection rate is a ratio between correctly labeled instances and total labeled instances. It is a percentage of positive predictions in specific class that are correct. It is defined by:

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP } + \text{FP}} \tag{15}$$

where TP and FP are the number of true positive and false positive predictions for the particular class.

The F-measure is the harmonic mean of precision and recall and it attempts to give a single measure of performance. A good classifier can provide both recall and precision values high. The F-measure is defined as:

$$F_{\beta} = \frac{(1+\beta)^{2} \cdot \mathrm{TP}}{(1+\beta)^{2} \cdot \mathrm{TP} + \beta^{2} \cdot \mathrm{FN} + \mathrm{FP}} \tag{16}$$

where β is the weighting factor. Here, $\beta = 1$, that is, precision and recall are equally weighted and used to measure the F_{β}-score which is also called as F1-measure.

The most common metric accuracy is defined as the ratio between sum of correct classifications and total number of classifications.

$$\text{Accuracy (A)} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{17}$$

8 Experimental Results

The experiments are carried out in MATLAB R2013a in Windows 7 operating system on computer with Intel Pentium i3 processor 2.10 GHz with 3 GB of RAM. As explained in feature extraction section, the six-dimensional features are extracted. The performance of the proposed feature method is tested on SVM, Naïve Bayes, and DTW classifiers to test the performance.

8.1 Results Obtained with SVM

The extracted features are fed to SVM with polynomial kernel. In polynomial kernel, different degrees (2, 3, 4, 5, and 6) were tested. Based on the classification results, degree 4 performs superior than the other kernel degrees. Further, it has been observed that increase in kernel degree does not give any improvement in performance.

The confusion matrix of the SVM with polynomial classifier on emotion dataset is shown in Table 4, where correct responses define the main diagonal, and most of the emotion classes like happy-walk, happy-sit, happy-jump, angry-sit, angry-jump, and fearful-walk are almost predicted well. An average recognition rate of SVM with polynomial kernel classifier on emotion dataset is 91.98 %. From this, some of the fearful-sit emotions are misclassified as happy-sit. Fearful-jump emotion is misclassified as happy-jump and angry-jump, respectively. Angry-walk is partly confused with happy-walk and fearful-walk, respectively.

Table 4 Confusion matrix for emotion and action recognition using SVM with polynomial classifier

Emotion/action	Happy (walk)	Happy (sit)	Happy (jump)	Angry (walk)	Angry (sit)	Angry (jump)	Fearful (walk)	Fearful (sit)	Fearful (jump)
Happy (walk)	**93.10**	0	0	6.90	0	0	0	0	0
Happy (sit)	0	**94.21**	0	0	2.51	0	3.28	0	0
Happy (jump)	0	0	**93.36**	0	0	6.64	0	0	0
Angry (walk)	4.21	0	0	**90.16**	0	0	5.63	0	0
Angry (sit)	0	6.79	0	0	**93.21**	0	0	0	0
Angry (jump)	0	0	8.58	0	0	**91.42**	0	0	0
Fearful (walk)	0	0	0	6.77	0	0	**93.23**	0	0
Fearful (sit)	0	10.68	0	0	0	0	0	**89.32**	0
Fearful (jump)	0	0	7.17	0	0	3.05	0	0	**89.78**

Table 5 Performance measure of the emotion-based actions using SVM with polynomial classifier

Emotion/action	Recall (%)	Precision (%)	F-measure (%)
Happy (walk)	93.10	95.67	94.37
Happy (sit)	94.21	84.36	89.01
Happy (jump)	93.36	85.57	89.29
Angry (walk)	90.16	86.83	88.47
Angry (sit)	93.21	97.38	95.25
Angry (jump)	91.42	90.42	90.92
Fearful (walk)	93.23	91.28	92.24
Fearful (sit)	89.32	100	94.36
Fearful (jump)	89.78	100	94.61
Average	91.98	92.39	92.06

Table 5 shows the average performance metrics of SVM with polynomial classifier; from the results, it is clearly indicated that proposed method gives higher recall = 91.98 %, precision = 92.39 %, and F-measure = 92.06 % (trade-off between precision and recall), where high recall value indicate that an SVM with polynomial classifier returned most of the relevant samples correctly.

8.2 Results Obtained with Naïve Bayes

The confusion matrix of the Naïve Bayes classifier on emotion dataset is shown in Table 6, where correct responses define the main diagonal, and most of the emotion classes like happy-sit, happy-jump, angry-sit, and fearful-walk are almost predicted well. An average recognition rate of Naïve Bayes classifier on emotion dataset is 89.91 %. From this, some of happy-walk and angry-walk emotions are misclassified as angry-walk and fearful-walk, respectively. Fearful-sit is mostly confused with happy-sit. In contrast, angry-jump and fearful-jump emotion are misclassified as happy-jump.

Table 7 shows the average performance metrics of Naïve Bayes classifier; from the results, it is evidently indicated that proposed method gives good recall = 89.91 %, precision = 90.16 %, and F-measure = 89.96 %. From the quantitative evaluation results, the proposed method has good recall, precision, F-measure, and accuracy for the Naïve Bayes classifier on emotion recognition dataset. It is found that the overall performance is dropped to 2 % on Naïve Bayes classifier, when compared to SVM with polynomial classifier.

Table 6 Confusion matrix for emotion and action recognition using Naïve Bayes classifier

Emotion/action	Happy (walk)	Happy (sit)	Happy (jump)	Angry (walk)	Angry (sit)	Angry (jump)	Fearful (walk)	Fearful (sit)	Fearful (jump)
Happy (walk)	**89.30**	0	0	8.10	0	0	2.60	0	0
Happy (sit)	0	**91.11**	0	0	4.21	0	2.61	2.07	0
Happy (jump)	0	0	**90.25**	0	0	8.26	0	0	1.49
Angry (walk)	4.81	0	0	**89.56**	0	0	5.63	0	0
Angry (Sit)	0	5.49	0	0	**91.41**	0	0	3.10	0
Angry (jump)	0	0	9.18	0	0	**89.36**	0	0	1.46
Fearful (walk)	2.39	0	0	7.42	0	0	**90.19**	0	0
Fearful (sit)	0	10.28	0	0	0	0	0	**89.72**	0
Fearful (jump)	0	0	8.53	0	0	3.21	0	0	**88.26**

Table 7 Performance measure of the emotion-based actions using Naïve Bayes classifier

Emotion/action	Recall (%)	Precision (%)	F-measure (%)
Happy (walk)	89.30	92.54	90.89
Happy (sit)	91.11	85.25	88.08
Happy (jump)	90.25	83.60	86.80
Angry (walk)	89.56	85.23	87.34
Angry (sit)	91.41	95.60	93.46
Angry (jump)	89.36	88.62	88.99
Fearful (walk)	90.19	89.27	89.73
Fearful (sit)	89.72	94.55	92.07
Fearful (jump)	88.26	96.77	92.32
Average	**89.91**	**90.16**	**89.96**

8.3 Results Obtained from DTW

The confusion matrix of the DTW classifier on emotion dataset is shown in Table 8, where correct responses define the main diagonal, and most of the emotion classes like happy-walk, happy-sit, happy-jump, angry-walk, angry-sit and fearful-walk are almost predicted well. An average recognition rate of DTW classifier on emotion dataset is 93.39 %. From this, some of fearful-sit emotions are misclassified as happy-sit. Angry-jump emotion is misclassified as happy-jump. Fearful-jump is mostly confused with happy-jump and fearful-jump, respectively.

Table 9 shows the average performance metrics of DTW classifier; from the results, it is clearly indicated that proposed method gives higher recall = 93.7 %, precision = 93.2 %, and F-measure = 93.4 %, where high recall value indicates that DTW classifier returned most of the relevant samples correctly. From the quantitative evaluation results, the proposed approach has a superior recall, precision, F-measure, and accuracy for the DTW classifier on emotion recognition dataset.

It is found that the overall performance is increased to 3.5 % on DTW classifier, when compared to SVM with polynomial and Naïve Bayes classifiers.

The potential use of automatic visual surveillance application is to detect and analyze abnormal situations. To achieve this, technological support and connected smart devices, which are becoming a reality, are used. The devices can monitor and can act as sensors to sense the environment and in particular to monitor human assistance and to ensure the public safety. They also process information, control traffic lights, lock doors, and remind people to take medications.

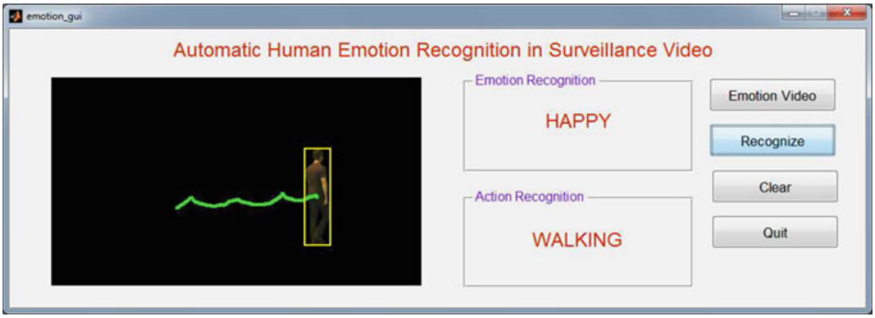
This chapter proposes an approach for emotion recognition system in the context of civil safety which raises an alarm when an abnormal emotion is detected. Figures 11, 12, and 13 shows the detection of happy-walking and angry-sit and fearful-jump while monitoring emotion of the person.

Table 8 Confusion matrix for emotion and action recognition using DTW classifier

Emotion/action	Happy (walk)	Happy (sit)	Happy (jump)	Angry (walk)	Angry (sit)	Angry (jump)	Fearful (walk)	Fearful (sit)	Fearful (jump)
Happy (walk)	**94.6**	0	0	5.41	0	0	0	0	0
Happy (sit)	0	**96.08**	0	0	1.96	0	1.96	0	0
Happy (jump)	0	0	**95**	0	0	5	0	0	0
Angry (walk)	2.56	0	0	**92.31**	0	0	5.13	0	0
Angry (sit)	0	5.71	0	0	**94.29**	0	0	0	0
Angry (jump)	0	0	9.38	0	0	**90.63**	0	0	0
Fearful (walk)	0	0	0	5	0	0	**95**	0	0
Fearful (sit)	0	10	0	0	0	0	0	**90**	0
Fearful (jump)	0	0	4.55	0	0	4.55	0	0	**90.91**

Table 9 Performance measure of the emotion-based actions using DTW classifier

Emotion/action	Recall (%)	Precision (%)	F-measure (%)
Happy (walk)	97.2	94.5	95.8
Happy (sit)	90.7	96	93.3
Happy (jump)	88.3	95	91.5
Angry (walk)	90	92.3	91.1
Angry (sit)	97	94.2	95.6
Angry (jump)	87.8	90.6	89.2
Fearful (walk)	92.6	95	93.8
Fearful (sit)	100	90	94.7
Fearful (jump)	100	90.9	95.2
Average	**93.7**	**93.2**	**93.4**



Automatic Human Emotion Recognition in Surveillance Video

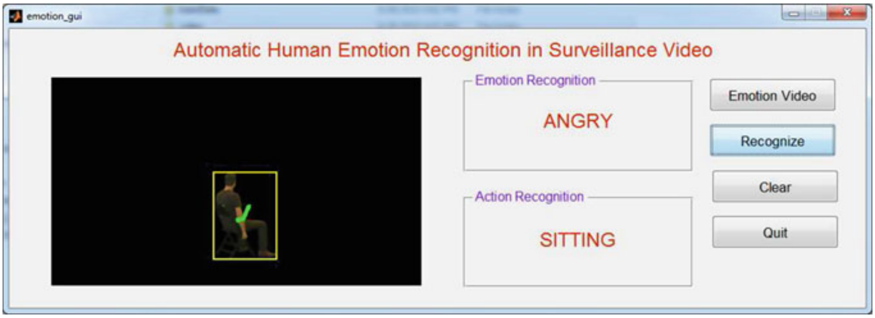
Emotion Recognition

HAPPY

Action Recognition

WALKING

Fig. 11 Snapshot of the happy emotion with walking action using DTW



Automatic Human Emotion Recognition in Surveillance Video

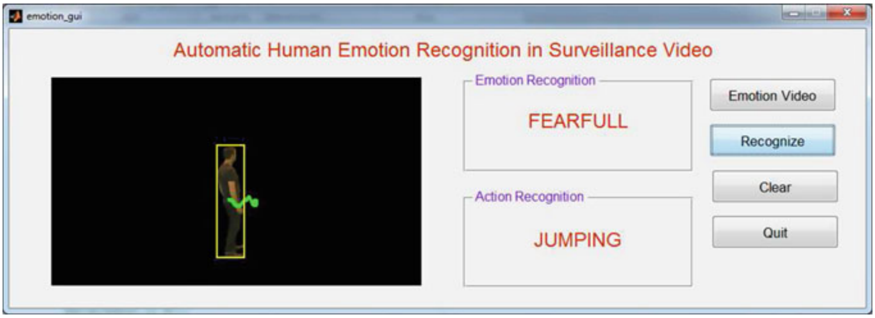
Emotion Recognition

ANGRY

Action Recognition

SITTING

Fig. 12 Snapshot of the angry emotion with sitting action using DTW


Automatic Human Emotion Recognition in Surveillance Video

Emotion Recognition

FEARFULL

Action Recognition

JUMPING

Fig. 13 Snapshot of the fearful emotion with jumping action using DTW

9 Conclusion

This chapter introduces a human emotion recognition using gesture dynamics in surveillance video. Experiments are conducted on emotion dataset considering different persons. The proposed gesture dynamics features are extracted from the emotion video sequences and modeled using SVM with polynomial, Naïve Bayes, and DTW; the performance measures such as recall, precision, F-measure, and accuracy were calculated. Experimental results show the overall accuracy of SVM with polynomial, Naïve Bayes, and DTW results as 91.98, 89.91, and 93.39 %, respectively. The performance results indicate that DTW outperforms SVM and Naïve Bayes classifiers. It is observed from the experiments that the system could not distinguish angry-jump, fearful-sit, and fearful-jump with high accuracy and is of future interest.

Smart screens kept at homes, offices, and shopping centers would provide information about the customer's interest about the products. Further, by dynamically recognizing emotions of buyers and their whereabouts, sellers could send them alerts related to their products, delivered to their mobile. Customer service could be improved without having customers to go through the terrific process of reporting problems to the sellers by recognizing their emotions.

Automatic emotion recognition is addressed in this chapter at the budding level. Lot more issues are to be addressed by the base problem itself, which attracts the researchers nowadays.

References

1. Castellano G, Villalba SD, Camurri A (2007) Recognising human emotions from body movement and gesture dynamics. Affective computing and intelligent interaction. Springer, Berlin, pp 71–82
2. Bernhardt D, Robinson P (2009) Detecting emotions from connected action sequences, visual informatics: bridging research and practice. Springer, Berlin, pp 1–11

3. Yoo H-W, Cho S-B (2007) Video scene retrieval with interactive genetic algorithm. Multimed Tools Appl 34(3):317–336
4. Ke S-R (2013) A review on video-based human activity recognition. Computers 2(2):88–131
5. Anagnostopoulos C-N, Iliou T, Giannoukos I (2015) Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artif Intell Rev 43(2):155–177
6. Rao KS, Koolagudi SG (2015) Recognition of emotions from video using acoustic and facial features. SIViP 9(5):1029–1045
7. Busso C (2004) Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th international conference on Multimodal interfaces. ACM
8. Gunes H, Piccardi M (2007) Bi-modal emotion recognition from expressive face and body gestures. J Netw Comput Appl 30(4):1334–1345
9. Karpouzis K et al (2007) Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. Artifical intelligence for human computing. Springer, Berlin, pp 91–112
10. Gross MM, Crane EA, Fredrickson BL (2010) Methodology for assessing bodily expression of emotion. J Nonverbal Behav 34(4):223–248
11. Hassan M et al (2014) A review on human actions recognition using vision based techniques. J Image Graph 2(1):28–32
12. Kessous L, Castellano G, Caridakis G (2010) Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. J Multimodal User Interfaces 3(1-2):33–48
13. Gunes H(2010) Automatic, dimensional and continuous emotion recognition
14. Kapoor A, Burleson W, Picard RW (2007) Automatic prediction of frustration. Int J Hum Comput Stud 65(8):724–736
15. Kapur A (2005) Gesture-based affective computing on motion capture data. Affective computing and intelligent interaction. Springer, Berlin, pp 1–7
16. Balomenos Themis et al (2004) Emotion analysis in man-machine interaction systems. Machine learning for multimodal interaction. Springer, Berlin, pp 318–328
17. Camurri A, Lagerlöf I, Volpe G (2003) Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. Int J Hum Comput Stud 59 (1):213–225
18. Yang Z, Ortega A, Narayanan S (2014) Gesture dynamics modeling for attitude analysis using graph based transform. In: 2014 IEEE international conference on image processing (ICIP). IEEE
19. Zaboleeva-Zotova AV (2013) Automated identification of human emotions by gestures and poses. In: 2013 BRICS congress on computational intelligence and 11th Brazilian congress on computational intelligence (BRICS-CCI & CBIC). IEEE
20. Cowie R, McKeown G, Douglas-Cowie E (2012) Tracing emotion: an overview. Int J Synth Emot 3(1):1–17
21. Alvandi EO (2011) Emotions and information processing: a theoretical approach. Int J Synth Emot 2(1):1–14
22. Salovey P, Mayer JD (1990) Emotional intelligence. Imagin Cogn Personal 9(3):185–211
23. Oker A et al (2015) A virtual reality study of help recognition and metacognition with an affective agent. Int J Synth Emot 6(1):60–73
24. Warwick K, Harrison I (2014) Feelings of a cyborg. Int J Synth Emot 5(2):1–6
25. Keefe Bruce D et al (2014) A database of whole-body action videos for the study of action, emotion, and untrustworthiness. Behav Res Methods 46(4):1042–1051
26. Müller M (2007) Dynamic time warping. Information retrieval for music and motion, pp 69–84
27. Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge
28. Mitchell TM, Michell T (1997) Machine learning. Mc-graw-Hill Series in Computer Science
29. Vapnik V (1998) Statistical learning theory