INTRODUCTION TO EMOTION RECOGNITION

AMIT KONAR AND ANISHA HALDER
Artificial Intelligence Laboratory, Department of Electronics and Telecommunication Engineering, Jadavpur University, Kolkata, India

ARUNA CHAKRABORTY
Department of Computer Science & Engineering, St. Thomas’ College of Engineering & Technology, Kolkata, India

A pattern represents a characteristic set of attributes of an object by which it can be distinguished from other objects. Pattern recognition aims at recognizing an object by its characteristic attributes. This chapter examines emotion recognition in the settings of pattern recognition problems. It begins with an overview of the well-known pattern recognition techniques, and gradually demonstrates the scope of their applications in emotion recognition with special emphasis on feature extraction, feature reduction, and classification. Main emphasis is given to feature selection by single and multiple modalities and classification by neural, fuzzy, and statistical pattern recognition techniques. The chapter also provides an overview of stimulus generation for arousal of emotion. Lastly, the chapter outlines the methods of performance analysis and validation issues in the context of emotion recognition.

1.1 BASICS OF PATTERN RECOGNITION

A pattern is a representative signature of an object by which we can recognize it easily. Pattern recognition refers to mapping of a set of patterns into one of several object classes. Occasionally, a pattern is represented by a vector containing the features of

INTRODUCTION TO EMOTION RECOGNITION

The pattern recognition process can be described by three fundamental steps, namely, feature extraction, feature selection, and classification. Figure 1.1 provides a general scheme for pattern recognition. The feature extraction process involves using one or more sensors to measure the representative features of an object. The feature selection module selects more fundamental features from a list of features. The classification module classifies the selected features into one of several object classes.

The pattern recognition problem can be broadly divided into two main heads: (i) supervised classification (or discrimination), and (ii) unsupervised clustering. In supervised classification, usually a set of training instances (or data points) comprising a set of measurements about each object along with its class is given. These data points with their class labels are used as exemplars in the classifier design. Given a data point with unknown class, the classifier once trained with the exemplary instances is able to determine the class label of the given data point. The classifier thus automatically maps an unknown data point to one of several classes using the background knowledge about the exemplary instances.

Beginners to the subject often are confronted with the question: how does the classifier automatically determine the class label of an unknown data point, which is not present in the exemplary instances. This is due to the inherent generalization characteristics of the supervised classifier.

In unsupervised classification, the class labels of the data points are not known. The learning system partitions the whole set of data points into (preferably) nonoverlapping subsets based on some measure of similarity of the data points under each subset. Each subset is called a class/cluster. Because of its inherent characteristics of grouping data points into clusters, unsupervised classification is also called clustering.

Both statistical decision theory and machine learning have been employed in the literature to design pattern recognition algorithms [1, 2]. Bayes’ theorem is the building stone of statistical classification algorithms. On the other hand there exists a vast literature on supervised and unsupervised learning [3], algorithms, which capture the inherent structural similarity [4] of the data points for application in pattern recognition problems.

EMOTION DETECTION AS A PATTERN RECOGNITION PROBLEM

Emotion represents the psychological state of the human mind and thought processes. Apparently, the process of arousal of emotion has a good resemblance with its manifestation as facial, vocal, and bodily gestures. This phenomenon has attracted researchers to determine the emotion of a subject from its manifestation. Although the
one-to-one correspondence from manifestation of emotion to a particular emotional state is yet to be proved, researchers presume the existence of such mapping to recognize the emotion of a subject from its manifestation.

Given the manifestation of an emotion, the task of recognizing the emotion, thus, is a pattern recognition problem. For example, facial expression–based emotion recognition requires extraction of a set of facial features from the facial expression of a given subject. Recognition of emotion here refers to classification of facial features into one of several emotion classes. Usually, a supervised classifier pretrained with emotional features as input and emotion class as output is used to determine the class of an unknown emotional manifestation.

Apparently, the emotional state of the human mind is expressed in different modes including facial, voice, gesture, posture, and biopotential signals. When a single mode of manifestation is used to recognize emotion, we call it a unimodal approach. Sometimes all modes are not sufficiently expressed. Naturally, recognition from a less expressed mode invites the scope of misclassification. This problem can be avoided by attempting to recognize an emotion from several modalities. Such a process is often referred to as multimodal emotion recognition.

1.3 FEATURE EXTRACTION

Feature extraction is one of the fundamental steps in emotion recognition. Features are obtained in different ways. On occasion features are preprocessed sensory readings. Preprocessing is required to filter noise from measurements. Sensory readings during the period of emotion arousal sometimes have a wide variance. Statistical estimates of the temporal readings, such as mean, variance, skewness, kurtosis, and the like, are usually taken to reduce the effect of temporal variations on measurements. Further, instead of directly using time/spatial domain measurements, frequency domain transforms are also used to extract frequency domain features. For example, frequency domain information is generally used for EEG (electroencephalogram) and voice signals. Frequency domain parameters are time invariant and less susceptible to noise. This attracted researchers to use frequency domain features instead of time domain.

Frequency domain features have one fundamental limitation in that they are unable to tag time with frequency components. Tagging the time with frequency contents of a signal is important, particularly for a certain class of signals, often labeled as nonstationary signals. EEG, for instance, is a nonstationary signal, the frequency contents of which change over time because of asynchronous firing of the neurons. Wavelet transform coefficients of an EEG signal represent time–frequency correlations and thus deserve to be one of the fundamental features for nonstationary EEG signals. We now briefly outline the features used in different modalities of emotion recognition.

1.3.1 Facial Expression–Based Features

The most common modality of emotion recognition is by facial expression analysis. Traditionally there exist two major classes of techniques for face/facial expression
representation and relevant feature extraction. The first one is called geometrical features. They rely on parameters of distinctive facial features such as eyes, mouth, and nose. On the other hand, appearance-based approach considers a face as an array of intensity values suitably preprocessed. This array is then compared with a face template using a suitable matrix.

### 1.3.1.1 Geometric Model–Based Feature Extraction

Deformable templates have been used for locating facial features. For example, Kass et al. [5] suggested the use of active contour models—snakes—for tracking lips in image sequences. They initialized snake on the lips in the facial image and showed that it is able to track lip movement accurately. However, it fails if there exists occlusion or other structure in the image.

Yuille et al. [6] employed deformable templates based on simple geometrical shapes for locating the eye and mouth. Yuille’s model incorporated shape constraints, but there is no proof that the form of a given model is sufficiently general to capture the deformable geometric shapes.

Researchers are taking a keen interest to represent geometric relations between facial information to extract facial features. In Reference 7, Craw, Tock, and Bennett, considered positional constraints in facial expressions to extract necessary features for emotion recognition.

Brunelli and Poggio considered a number of high dimensional [8] measurements or location of a number of key points in a single image or an image sequence for facial image interpretation.

Kirby and Sirovich [9] took attempts to decompose facial image into a weighted sum of basis images or eigen faces using Karhunen and Loeve expansion. They considered 50 expansion coefficients and were able to reconstruct an approximation of the facial expansion using these parameters.

### 1.3.1.2 Appearance-Based Approach to Feature Extraction

Appearance-based approach involves preprocessing followed by a compact coding through statistical redundancy reduction. The preprocessing in most cases is required to align the geometry in face image, for instance, by having the two eyes and nose tip at fixed positions through affine texture warping [10]. Optical flow or Gabor wavelets are used to capture facial appearance motion and robust registration, respectively, for successful recognition.

Pixel-based appearance is often represented by a compact coding. Usually statistical reduction principle is used to represent this coding. The unsupervised learning techniques used for compact coding include Principal Component Analysis (PCA), Independent Component Analysis (ICA), Kernel-PCA (KPCA), local feature analysis, and probability density estimation. Supervised learning techniques including Linear Discriminant Analysis (LDA) and Kernel Discriminant Analysis (KDA) are also used for compact coding representation.

The main drawback of PCA-based compact coding is that it retains some unwanted variations. It is also incapable of extracting local features that offer robustness against changes in local region or occlusions. ICA produces basis vectors that are more spatially local than those of PCA. Thus, ICA is sensitive to occlusion and pose
FEATURE EXTRACTION

variations. ICA retains higher order statistics and maximizes the degree of statistical independence of features [11, 12].

Recently Scholkopf et al. [13] extended the conventional PCA to KPCA, which is able to extract nonlinear features [14]. However, like PCA, KPCA captures the overall variance of all patterns and is not necessarily optimal for discrimination.

Statistical supervised learning such as LDA attempts to find the basis vectors maximizing the interclass distance and minimizing the intraclass distance. Similarly, KDA determines the most significant nonlinear basis vector to maximize the interclass distance while minimizing the intraclass distance. Among the other interesting works, the following need special mentioning.

Cohn et al. [15] proposed a facial action recognition technique that employs discriminant function analysis of individual facial regions, including eyebrows, eyes, and mouth. They used two discriminant functions for three facial actions of the eyebrow region, two discriminant functions for three facial actions of the eye region, and five discriminant functions for nine facial actions of the nose and mouth region. The classification accuracy for the eyebrow, eye and nose, mouth regions are 92, 88, and 88%, respectively.

In Reference 16 Essa and Pentland proposed a novel control-theoretic method to extract the spatiotemporal motion–energy representation of facial motion in an observed expression. They generated the spatiotemporal templates for six different facial expressions, considering two facial actions, including smile and raised eyebrows for two subjects. Templates are formed by averaging the patterns of motion generated by two subjects exhibiting a certain expression. The Euclidean norm of the difference between the motion–energy template and the observed motion energy is defined as a metric of similarity of the motion energies. A recognition accuracy of 98% is achieved while experimenting with 52 frontal-view image sequences of eight people having distinct expressions.

Kimura and Yachida in Reference 17 modeled facial images by a potential net and attempted to fit the net to each frame of a facial image sequence. The deformed version of the potential net is used to match the expressionless face, typically the first frame of the sequence. The variation in the nodes of the deformed net is used for subsequent processing. In their own experiments, the authors in Reference 17 considered a six image sequence of emotional expressions experienced by a subject with gradual variation in the strength of expressions from expressionless (relaxed) to a maximum expression. The experiment was repeated for three emotions: anger, happiness, and surprise. PCA has been employed here to classify three emotions using standard eigen space analysis.

In Reference 18 Lucey et al. detected pain from the movement of facial muscles into a series of action units (AUs), based on the Facial Action Coding System (FACS) [19]. For this novel task, they considered three types of Active Appearance Model (AAM) features: (i) similarity-normalized shape features (SPTS), (ii) similarity-normalized appearance features (SAPP), and (iii) canonical-normalized appearance features (CAPP). AAM features are used here to track the face and to extract visual features, based on facial expressions using the FACS. They obtained classification accuracy of 75.1, 76.9, and 80.9% using SAPP, SPTS, and CAPP,
INTRODUCTION TO EMOTION RECOGNITION

respectively, using Support Vector Machine (SVM) classifier first and then improving the performance (Fusion of Scores) by linear logistical regression (LLR).

Tian et al. in Reference 20 proposed a new method for recognizing AUs for facial expression analysis. They used both permanent and transient features for their work. Movement of eyebrow, cheek, eyes, and mouth are considered permanent features. On the other hand, deeping of facial furrows is considered a transient features. They used different feature extraction algorithms for different features. For the lips, they used lip-tracking algorithm. For eyes, eyebrows, and cheeks, they considered Lucas–Kanade algorithm, and for the transient features they employed Canny edge detector algorithm. Two neural network (NN) based classifiers are considered to recognize the changes in AUs: one for six upper-face AUs and the other for ten lower-face AUs of the FACS. A percentage accuracy of 95.4% is obtained for upper-face AUs and 95.6% for lower-face AUs.

In Reference 21 Kim and Bien designed a personalized classifier from facial expressions using soft computing techniques. They used degree of mouth openness \( f_1 \), degree of eye openness \( f_2 \), the vertical distance between the eyebrows and the eyes \( f_3 \), degree of nasolabial root wrinkles (NLR) \( f_4 \), and degree of nasolabial furrows (NLF) \( f_5 \) in their classifier design. These features are extracted from facial expressions by different techniques. For example, \( f_1 \) and \( f_2 \) are extracted by a human visual system based approach. \( f_1 \) is measured by combining global features (the height ratio and the area ratio between the whole face and the mouth region) and a local feature (Gabor–Gaussian feature). For \( f_2 \) they used the "dip" feature in log-polar mapped image and the Gabor-filter coefficients. For \( f_3, f_4, \) and \( f_5 \), which are transient components, they used the Euclidean distance \( f_4 \) and the Gabor-filtered coefficients \( f_5 \). Image features are extracted from four sets of facial expression data to show effectiveness of the proposed method, which confirms considerable enhancement of the whole performance by using Fuzzy Neural Nets (FNN) based classifier.

Huang et al. in Reference 22 proposed a novel approach to recognize facial expression using skin wrinkles. They considered many features like eyes, mouth, eyebrows, nostrils, nasolabial folds, eye pouches, dimples, forehead, and chin furrows for their research and used Deformable Template Model (DTM) and Active Wavelet Network (AWN) for extracting those features. Classification accuracy obtained by using Principle Component Analysis and Neural Network is around 70%.

In Reference 23, Kobayashi and Hara recognized basic facial expressions by using 60 facial characteristic points (FCP) from three components of the face (eyebrows, eyes, and mouth). These features are extracted by manual calculation and emotions are classified by neural network.

In Reference 24 a real-time automated system was modeled by Anderson and McOwan, for recognition of human facial expressions. Here, the muscle movements of the human face are considered as features after tracking the face. A modification of spatial ratio template tracker algorithm is used here for tracking the face first and later to determine the motion of the face by optical flow algorithm. A percentage accuracy of 81.82% has been obtained by using SVM as classifier.

Otsuka and Ohya [25] considered a matching of temporal sequence of the 15D feature vector to the models of the six basic facial expressions by using a specialized
Hidden Markov Model (HMM). The proposed HMM comprises five states, namely, relaxed (S1, S5), contracted (S2), apex (S3), and relaxing (S4). The recognition of a single image sequence here is realized by considering transition from the final state to the initial state. Further, the recognition of multiple sequences is accomplished by considering transitions from a given final state to initial states of all feasible categories. The state-transition probability and output probability of each state are obtained from sampled data by employing Baum–Welch algorithm. The k-means clustering algorithm here has been used to estimate the initial probabilities. The method was tested on the same subjects for whom data was captured. Consequently, the feasibility of the proposed technique for an unknown subject is questionable. Although the proposed method was labeled as good, no justification was given in favor of its goodness. Besides the above, the works of Ekman [27–38], Pantic [39–46], [48–51], Cohn [52–57], Konar [58–69] and some others [70–73] deserve special mention.

1.3.2 Voice Features

Voice features used for emotion recognition include prosodic and spectral features. Prosodic features are derived from pitch, intensity, and first formant frequency profiles as well as voice quality measures. Spectral features include Mel-frequency cepstral coefficients (MFCC), linear prediction cepstral coefficients (LPC), log frequency power coefficients (LFPC), perceptual linear prediction (PLP) coefficients. We now briefly provide an overview of the voice features below.

1. Pitch represents the perceived fundamental frequency of a sound. Fundamental frequency is defined as the frequency at which the vocal cords vibrate during speech.

2. A formant is a peak in a frequency spectrum that results from the resonant frequencies of any acoustical system. For human voice, formants are recognized as the resonance frequencies of the vocal tracts. Formant regions are not directly related to the fundamental frequency and may remain more or less constant as the fundamental changes. If the fundamental is low in the formant range, the quality of the sound is rich, but if the fundamental is above the formant regions, the sound is thin. The first three formants: $F_1$, $F_2$, and $F_3$ are more often used to disambiguate the speech.

3. Power spectral density describes the distribution of power of a speech signal with frequency and also shows the strength (signal energy is strong or weak at different frequency) of the signal as a function of frequency. The energy or power (average energy per frame) in a formant comes from the sound source (vibration of the vocal folds, frequency of the vocal tract, movement of lips and jaw). The energy in the speech signal $x(n)$ is computed as

$$E_x = \sum_{n=0}^{N-1} x^2(n)$$  (1.1)
8 INTRODUCTION TO EMOTION RECOGNITION

The power of the signal $x(n)$ is the average energy per frame:

$$P_x = \frac{1}{N} \sum_{n=0}^{N-1} x^2(n) \quad (1.2)$$

where $N$ is total no. of samples in a frame.

4. “Jitter is defined as perturbations of the glottal source signal that occurs during vowel phonation and affect the glottal pitch period” [75]. Let $u[n]$ be a pitch period sequence. Then we define absolute jitter by

$$\frac{1}{N+1} \sum_{n=0}^{N-1} |u(n+1) - u(n)|$$

5. “Shimmer is defined as perturbations of the glottal source signal that occur during vowel phonation and affect the glottal energy” [75]. Let $u[n]$ be a peak amplitude sequence of $N$ samples. Then absolute shimmer is given by

$$\frac{1}{N+1} \sum_{n=0}^{N-1} |u(n+1) - u(n)|$$

6. MFCC [76] is a widely used term in speech and speaker recognition. However, the definition of MFCC requires defining two important parameters: “Mel scale” and Mel-frequency spectrum. The Mel scale is defined as

$$f_{mel} = 1125 \ln(1 + f/700)$$

where $f$ is the actual frequency in Hz. Mel-frequency cepstrum (MFC) is one form of representation of the short-term power spectrum of sound, based on a linear cosine transformation of a log-power spectrum on a nonlinear “mel” scale of frequency.

Mel-frequency cepstral coefficients are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear “spectrum-of-a-spectrum”). The difference between the cepstrum and the Mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the Mel scale, which approximates the human auditory system’s response more closely than the linearly spaced frequency bands used in the normal cepstrum.

7. A speech sample can be modeled as a linear combination of its past samples. A unique set of predictor coefficients is determined by minimizing the sum of the squared differences between the actual speech samples and the linearly predicted ones. These predictor coefficients are referred to as linear prediction-based cepstral coefficient (LPCC) [77].
Busso et al. in Reference 78, presented a novel approach for emotion detection from emotionally salient aspects of the fundamental frequency in the speech signal. They selected pitch contour (mean, standard deviation, maximum, and minimum range of sentence- and voice-level features of pitch) as features for their experiment. Pitches obtained from emotional and neutral speech are compared first by symmetric KLD (Kullback–Leibler Distance). Then pitch features are quantified by comparing nested Logistic Regression Models. They used GMM (Gaussian Mixture Model) and LDC (Linear Discriminant Classifier) for classification process and obtained accuracy over 77%.

In Reference 79, Lee et al. detected emotions in spoken dialogues. Features, they used in the paper are pitch, formant frequencies, energy, and timing features like speech duration rate, ratio of duration of voiced and unvoiced region, and duration of the longest voiced speech. In this paper, irrelevant features are eliminated from the base feature set by forward selection (FS) method, and then a feature set is calculated by PCA. This novel approach improved emotion classification by 40.7% for males and 36.4% for females using LDC and k-NN (k-Nearest Neighborhood classifier) for emotion classification.

Wu et al. [80] proposed a new method for emotion recognition of affective speech based on multiple classifiers using acoustic–prosodic information and semantic labels. Among the acoustic–prosodic features, they selected pitch, intensity, formants and formant bandwidth, jitter-related features, shimmer-related features, harmonicity-related features, and MFCC. They derived Semantic Labels from HowNet (Chinese Knowledge Base) to extract EAR (Emotion Association Rules) from the recognized word sequence of the affective speech. They used multiple classifiers like GMM, SVM, MLP, (Multilayer Perceptron), MDT (Meta Decision Tree), and Maximum Entropy Model (MaxEnt) and got an overall accuracy of 85.79%.

Kim et al. [81] have developed an improved emotion recognition scheme with a novel speaker-independent feature. They employed orthogonal–linear discriminant analysis (OLDA) for extracting speech features, that is, ratio of a spectral flatness measure (SFM) to a spectral center (RSS), pitch, energy, and MFCC. They used GMM as a classifier for emotion recognition. An average recognition rate of 57.2% (±5.7%) at a 90% confidence interval can be obtained by their experiment. Among the other research works on speech, the work of Mower [82–86], Narayanan [78–94], Wu [95–98], Schuller [99–116], and some others [117–133] deserve special mention.

### 1.3.3 EEG Features Used for Emotion Recognition

Electroencephalogram is an interesting modality for emotion recognition. Under a hostile environment, people sometimes attempt to conceal the manifestation of their emotional states in facial expression and voice. EEG, on the other hand, gives a more realistic modality of emotion recognition, particularly, due to its temporal changes during arousal of emotion, and thus, concealment of the emotion in EEG is not feasible.

Usually the frontal lobe of the human brain is responsible for cognitive and emotion processing. There exists an internationally accepted 10-20 system for electrode placement on the scalp. Such placement of electrodes ensures that most of the
INTRODUCTION TO EMOTION RECOGNITION

Brain functions, such as motor activation, emotion processing, reasoning, etc., can be retrieved correctly from the EEG signals obtained from these channels. In the 10-20 system (shown in Figure 1.2) of electrode placement, the channels $F_3$, $F_4$, $F_{p1}$, and $F_{p2}$ are commonly used for emotion recognition.

Both time- and frequency-domain parameters of EEG are used as features for the emotion classification problems. Among the time-domain features, adaptive auto-regressive (AAR) and Hzorth parameters, and among the frequency-domain features power-spectral density is most popular. EEG being a nonstationary signal, its frequency contents change widely over time. Time–frequency correlated features thus carry essential information of the EEG signal. Wavelet transform coefficients are important examples of time–frequency correlated features. In our study [134, 135], we considered wavelet coefficients, power spectral density, and also AAR parameters [136] for feature extraction. Typically the length of such feature vectors is excessively high, and thus, a feature reduction technique is employed to reduce the length of vectors without losing essential features.

Petrantonakis et al. [137] proposed a novel approach to recognize emotion from brain signals using a novel filtering procedure, namely, hybrid adaptive filtering (HAF) and higher order crossings (HOC) analysis. HAF was introduced for an efficient extraction of the emotion-related EEG characteristics, developed by applying Genetic Algorithms (GA) to the Empirical Mode Decomposition (EMD) based representation of EEG signals. HOC analysis was employed for feature extraction from the HAF-filtered signals. They introduced a user-independent EEG-based emotion recognition system for the classification of six typical emotions, including happiness, surprise, anger, fear, disgust, and sadness. The EEG signals were acquired from $F_{p1}$, $F_{p2}$, $F_3$, and $F_4$ positions, according to the 10-20 system from 16 healthy subjects using three EEG channels through a series of facial-expression image projections, as a Mirror Neuron System based emotion elicitation process. For an extensive evaluation of the classification performance of the HAF–HOC scheme, Quadratic Discriminant Analysis (QDA), $k$-Nearest Neighbor (k-NN), Mahalanobis Distance (MD), and Support Vector Machines (SVMs) were adopted. For the individual-channel case, the best
results were obtained by the QDA (77.66% mean classification rate), whereas for the combined-channel case, the best results were obtained using SVM (85.17% mean classification rate).

In Reference 138, Petrantonakis et al. proposed a novel method for evaluating the emotion elicitation procedures in an EEG-based emotion recognition setup. By employing the frontal brain asymmetry theory, an index, namely Asymmetry Index (AsI), is introduced, in order to evaluate this asymmetry. This is accomplished by a multidimensional directed information analysis between different EEG sites from the two opposite brain hemispheres. The proposed approach was applied to three-channel ($F_{p1}$, $F_{p2}$, and $F_3/F_4$ 10/20 sites) EEG recordings drawn from 16 healthy right-handed subjects. For the evaluation of the efficiency of the AsI, an extensive classification process was conducted using two feature-vector extraction techniques and an SVM classifier for six different classification scenarios in the valence/arousal space. This resulted in classification results up to 62.58% for the user-independent case and 94.40% for the user-dependent one, confirming the efficacy of AsI as an index for the emotion elicitation evaluation.

Yuan-Pin Lin et al. [139] developed a new idea to recognize emotion from EEG signals while listening to music. In this study, EEG data were collected through a 32-channel EEG module, arranged according to the international 10-20 system. Sixteen excerpts from Oscar’s film soundtracks were selected as stimuli, according to the consensus reported from hundreds of subjects. EEG signals were acquired from 30 channels. The features selected include power spectrum density (PSD) of all the 30 channels. SVM was successfully employed to classify four emotional states (joy, anger, sadness, and pleasure) using the measured PSDs. The best result for classification accuracy obtained is found to have mean 82.29% with a variance of 3.06% using a 10 times of 10-fold cross-validation scheme across 26 subjects. A few more interesting works on EEG-based emotion recognition that need special mention include References 140–153.

1.3.4 Gesture- and Posture-Based Emotional Features

Gestures are expressive and meaningful motions, involving hands, face, head, shoulders, and/or the complete human body. Gesture recognition has a wide range of applications, such as sign language for communication among the disabled, lie detection, monitoring emotional states or stress levels of subjects, and navigating and/or manipulating in virtual environments.

Recognition of emotion from gestures is challenging as there is no generic notion to represent a subject’s emotional states by her gestures. Further, the gestural pattern has a wider variation depending on the subject’s geographical origin, culture, and the power and intensity of his or her expressions.

Gestures can be static, considering a single pose or dynamic with a prestroke, stroke, and poststroke phases [154]. Automatic recognition of continuous gestures requires temporal segmentation. The start and end points of a continuous gesture are often useful to segregate it from the rest. Segmentation of a gesture sometimes is difficult as the preceding and the following gestures often are similar.
INTRODUCTION TO EMOTION RECOGNITION

The most common gestural pattern, often used in emotion recognition, is the hand movements. Glowinski et al. [155] proposed an interesting technique for hand (and head) gesture analysis for emotion recognition. They considered a “bounding” triangle formed by the centroids of the head and hands, and determined several parameters of the 3D triangle to extract the features of individual hand gesture representative of emotions. A set of triangles obtained from a motion cue is analyzed to extract a large feature vector, the dimension of which is reduced later by PCA. A classification technique is used to classify emotion from a reduced 4D data space. Methodologies of feature reduction and classification will be discussed in a subsequent section later.

Camurri et al. [156] classified expressive gestures from the human full body movement during the performance of the subject in a dance. They identified motion cues and measured overall duration, contraction index, quantity of motion, and motion fluency. On the basis of these motion cues, they designed an automated classifier to classify four emotions (anger, fear, grief, and joy).

Castellano et al. [157] employed hand gestures for emotion recognition. They considered five different expressive motion cues, such as quantity of motion and contraction index of the body (degree of contraction and expansion of the body), velocity, acceleration, and fluidity (uniformity of motion) using Expressive Gesture Processing Library [158], and determined emotion from the cues by direct classification of time series.

In Reference 154, the authors considered both static and dynamic gesture recognition. While static gestures can be recognized by template matching, dynamic gesture is represented as a collection of time-staggered states, and thus can be modeled with HMM, sequential state machines (SSM), and discrete time differential neural nets (DTDNN). Preprocessing of the dynamic gestures includes tracking of important points or regions of interest in the image frames describing temporal states of the gesture. The tracking is often performed using particle filtering and level sets. Although particle filters and level sets have commonality in the usage of tracking in images, the principles used therein have significant difference. Usually a particle filter tracks a geometric-shaped region, circle or rectangle, on the image. So, in a nonrigid video, if the points enclosed in a region (of an image frame) have different directions of velocity, all the points of the reference frame cannot be tracked in subsequent frames. However, in a level set, the points enclosed in a region (of a frame) can be tracked by a nonlinear boundary, which may change its shape over the subsequent frames to keep track of all the points in the region of interest. Both the methods referred to above can track the hand and head gestures. For other important works on emotion recognition using gesture features, readers may consult the references aside 159–169.

1.3.5 Multimodal Features

Multimodality refers to analysis of different manifestations of emotion, including facial expression, voice, brain signals, body gesture, and physiological reactions. A few well-known multimodal schemes for emotion recognition are outlined below.
1.3.5.1 AudioVisual Zheng et al. [170] proposed an interesting approach to audiovisual emotion recognition. They considered 12 predefined motions of facial features, called Motion Units (MU) and 20 prosodic features, including pitch, RMS energy, formants F1–F4 and their bandwidths, and all of their corresponding derivatives. They achieved a person-independent classification accuracy of 72.42% using multi-stream HMM.

Mower et al. [171] designed a scheme for audiovisual emotion recognition. They considered both distance features between selective facial points and prosodic/spectral features of speech to recognize the emotion of an unknown subject. The distance features are selected within and between two of the following four regions: cheek, mouth, forehead and eyebrow. For example, the metrics considered refer to the distance from top of the check to the eyebrow, lower part of the cheek to mouth/nose/chin, relative distance features between pairs of selected points on cheek, and average positional features. Similarly they considered mouth features, such as mouth opening/closing, lips puckering, and distance of lip corner and top from the nose. The prosodic features extracted from speech include pitch and energy, and the spectral features employed include MFCC. They employed Emotion-Profile Support Vector Machines (EP-SVM) to obtain classification accuracy of 68.2%.

Busso et al. in Reference 172, proposed a new method to recognize emotion using facial expressions, speech, and multimodal information. They considered two approaches indicating fusion of two modalities at decision and feature levels. They employed 102 markers on the given facial image to determine the motion and alignment of the marked data points during utterance of 258 sentences expressing the emotions. They also considered prosodic features, such as pitch and intensity and reduced the feature dimensions by PCA. They classified the emotions by considering both individual visual and voice features, and obtained a classification accuracy of 85% for face data, 70.9% for voice data, and 89.1% for bimodal (face plus voice together) information.

1.3.5.2 Facial Expression–Body Gesture In a recent paper, Gunes and Picardi [173] consider automatic temporal segment detection and affect recognition from facial and bodily manifestation of emotional arousal. The main emphasis of the paper lies in the following thematic study. First, they demonstrate through experiments that affective faces and bodily gestures need not be strictly synchronous, although apparently they seem to occur jointly. Second, they observed that explicit detection of the temporal phases improves the accuracy of affect recognition. Third, experimental results obtained by them reveal that multimodal information including facial expression and body gesture together perform a better recognition of affect than only facial or body gestures. Last, they noticed that synchronous feature-level fusion achieves better performance than decision-level fusion.

1.3.5.3 Facial Expression–Voice–Body Gesture Nicolaou, Gunes, and Pantic, in Reference 174, have taken 20 facial feature points as facial features, MFCC, energy, RMS energy, pitch as speech features, and 5 shoulder points as body gesture feature for Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities
in Valence-Arousal Space. They introduced Bidirectional Long Short-Term Memory Neural Networks (BLSTM-NN) and Support Vector Regression (SVR) classifier for emotion classification and concluded that BLSTM-NN gives better performance than SVR.

Castellano et al. in Reference 175 introduced a novel approach to emotion recognition using multiple modalities, including face, body gesture, and speech. They selected 19 facial feature points as facial features, MFCC, pitch values, and lengths of voiced segments as speech features, and 80 motion features for each gesture as body gesture features. They trained and tested a model with a Bayesian classifier, using a multimodal corpus with 8 emotions and 10 subjects. To fuse facial expressions, gestures, and speech information, two different approaches were implemented: feature-level fusion, where a single classifier with features of the three modalities is used; and decision-level fusion, where a separate classifier is used for each modality and the outputs are combined a posteriori. Lastly, they concluded that the fusion performed at the feature level provided better results than the one performed at the decision level.

**1.3.5.4 EEG–Facial Expression** Chakraborty et al. [176] correlated stimulated emotion extracted from EEG and facial expression using facial features, including eye-opening, mouth-opening, and eyebrow constriction and EEG features, including frequency domain, time domain, and spatiotemporal features. They considered frequency domain features, such as peak power and average powers of $\alpha$, $\beta$, $\gamma$, $\theta$, and $\delta$ bands, time domain features, including 16 Kalman filters coefficients, and spatiotemporal features including 132 wavelet coefficients. They employed a feed-forward neural network to train it with a set of experimental instances using the well-known Back-propagation algorithm. The resulting network on convergence is capable of classifying instances to a level of 95.2% classification accuracy.

**1.3.5.5 Physiology** In Reference 177, Picard et al. proposed that for developing a machine’s ability to recognize human affective state, machines are expected to possess emotional intelligence. They performed an experiment considering four physiological signals: electromyogram (EMG), blood pressure volume, Hall effect, and respiration rate, taken from four sensors. One additional physiological signal, heart rate (H) here has been calculated as a function of the inter-beat intervals of the blood volume pressure. In their analysis, they used a combination of sequential forward search (SFFS) and Fisher projection (FP), called SFFS-FP, for selecting and transforming the features. A classification accuracy of 81% was obtained by using maximum a posteriori (MAP) classifier for SFFS-FP.

**1.3.5.6 Facial Expression–Voice–Physiology** Soleymani et al. [178] introduced a multimodal database for affect recognition and implicit tagging. They chose 27 subjects and recorded their videos of facial and bodily responses while watching 20 emotional videos. Features used for their experiment include distance metrics of the eye, eyebrow, and mouth as facial features, audio and vocal expressions, eye
gaze, pupil size, electrocardiograph (ECG), galvanic skin response (GSR), respiration amplitude, and skin temperature as physiological features. The main contribution of this research lies in the development of a large database of recorded modalities with high qualitative synchronization between them making it valuable to the ongoing development and benchmarking of emotion-related algorithms. The resulting database would provide support to a wide range of research on emotional intelligence, including data fusion, synchronization studies of modality, and many others.

1.3.5.7 EEG–Physiological Signals  Takahashi, in Reference 179, undertook an interesting research on emotion recognition from multimodal features including EEG, Pulse, and Skin Conductance. They collected psychological data of 12 subjects. The experimental setup contains a set of three sensors and two personal computers; one PC being used to present stimulus to a subject, while the other is used to acquire biopotential signals stated above. They used SVM classifier for emotion recognition from biopotential signals and acquired a classification accuracy of 41.7% for five emotions including joy, anger, sadness, fear, and relaxation. Among the other works on emotion recognition References 180–188 deserve special mention.

1.4 FEATURE REDUCTION TECHNIQUES

EEG and voice features usually have a high dimensionality and many of the experimentally obtained features are not independent. The speed of classifiers is often detrimental to the dimension of features. Feature reduction algorithms are required to reduce the dimensionality of features. Both linear and nonlinear feature reduction techniques are employed for emotion recognition.

Linear reduction techniques employ the characteristics of real symmetric matrices to extract independent features. In other words, the eigen vectors of real symmetric matrices are orthogonal (independent) to each other. Further, the larger eigen values of a system carry more information than the others. So, the eigen vectors corresponding to the large eigen values are used to reduce data dimensionality of a given linear system.

Linear reduction principles have gained popularity for their simplicity in use. However, on occasion researchers prefer nonlinear reduction techniques to their linear counterparts to improve precision and reliability of the classifier. Among the nonlinear feature reduction techniques, the most popular is rough set–based feature reduction. In this section, we briefly outline a few well-known linear and nonlinear feature reduction techniques.

1.4.1 Principal Component Analysis

Principal Component Analysis is one of the most popular linear feature reduction techniques. PCA represents $N$ measurements for $M$ subjects as a $M \times N$ matrix $A$, and computes $A^T A$ to obtain a real symmetric matrix $B$ of dimension $N \times N$. Now the $N$ eigen values of $B$ are evaluated, and the results are sorted in descending order.
INTRODUCTION TO EMOTION RECOGNITION

It is known that the larger eigen values have higher contribution in representing system characteristics, and thus to reduce features, we take \( k \) no. of eigen vectors corresponding to the first \( k \) eigen values of the list. The eigen vectors are arranged in columns, and the matrix is called Eigen Vector (EV) matrix of \((N \times k)\) dimension. Now, for each measurement vector \( a_i \), taken from the \( i \)th row of the matrix \( A \), we take a projection of \( a_i \) on the eigen space by multiplying \( A \) by EV matrix, and thus obtain \( a'_i \), where \( a'_i \) has dimension \((1 \times k)\). This is repeated for all \( i = 1 \) to \( N \), and thus the feature vectors \( a_i \) are mapped to \( k \)-dimensional vectors, where \( k \ll N \).

PCA is good for feature reduction of linear systems. If PCA is used for feature reduction of nonlinear systems, where the functional relationship between any two features is nonlinear, PCA sometimes loses important information. Researchers use PCA for its high efficiency and accuracy. Some important work using PCA include References 17, 79, 87, and 189–193.

1.4.2 Independent Component Analysis

Independent Component Analysis is a good choice to separate the sources from mixed signals. Particularly, in EEG and ECG the time series data \( x_t \) at time \( t \) is a nonlinear function \( f(.) \) of previous time samples, that is, \( x_t = f(x_0, x_1, \ldots, x_{t-1}) \). EEG signals taken from the forehead of a subject are often contaminated with eye-blinking signals, called electrooculogram (EOG). Further, the signal obtained at an electrode located on the scalp/forehead is due to the contribution of a number of signal sources at the neighborhood of the electrode. Elimination of EOG from EEG data and identification of the source signals can be performed together by ICA. One fundamental (although logical) restriction of using ICA lies in the inequality: \( C > S \), where \( C \) represents the number of EEG channels, and \( S \) stands for the number of independent signal sources. ICA has been widely used in the literature [194–198] to recognize emotion from facial expressions.

1.4.3 Evolutionary Approach to Nonlinear Feature Reduction

Evolutionary algorithms are population-based meta-heuristic optimization algorithms, which rest on the Darwinian principle of the survival of the fittest. The primary aim of this class of algorithms is to determine near optimal solutions, if not global, from a set of trial solutions through an evolutionary process determined by a set of operators like crossover, mutation, and selection. The most popular member of this class is Genetic Algorithm [199–201], devised by Prof. Holland approximately a half century ago. Among other members Differential Evolution (DE) [202, 203] is most popular for its structural and coding simplicity and exceptional performance in optimization problems.

Given a set of feature vectors (also called data points) and class labels for each vector, to implement the evolutionary feature reduction, we first use a supervised learning–based classifier to classify the data points into a fixed number of classes \( c \). Next we reduce the dimension of data points by dropping one feature randomly at a time, and again classify the data points into \( c \) number of classes. If the resulting
classes do not differ significantly with the previous classes, then the dropped feature has no major significance. The GA or DE is used to randomly select $k$ number of features at a time, and classify the data points into $c$ classes, and test whether the classes generated have significant difference with the classes obtained from the original dataset after classification. Since $k$ is randomly selected in $[1, n]$, where $n$ is the total number of features in the original dataset, we at the end of the search process expect to find a suitable value of $k$, for which the classes would be similar with the classes of the original dataset. Thus high dimensional features are reduced to $k$-dimension for $k \ll n$.

1.5 EMOTION CLASSIFICATION

This section provides principles of several approaches to emotion classification by pattern recognition techniques.

1.5.1 Neural Classifier

Neural networks have widely been used in emotion classification by facial expressions and voice. Both supervised and unsupervised neural architectures are employed in emotion classifiers. The supervised neural networks require a set of training instances. During the training process, the network encodes the connection weights in a manner, such that for all the input components of the training instances, the network can reproduce the output components of the corresponding training instances correctly as listed in the training instances. After the encoding is completed, the trained network can be used for testing. In the testing phase, an unknown input instance is submitted to the network, and the network generates the output instance using the encoded weights. In case of emotion recognition, the output of the neural net usually represents emotion classes, whereas input of the neural net represents a set of features extracted from facial expression/voice/gesture of the subjects. Naturally, a neural network pretrained with emotional features as the input and emotion classes as the output would be able to classify a specific emotional expression into one of several emotion classes.

1.5.1.1 Back-Propagation Algorithm

Among the well-known neural topologies, the Back-propagation is most common. Weight adaptation in the Back-propagation neural net is performed by Newtonian Gradient/Steepest descent learning principle. Let $W_{ij}$ be the connection weight between neuron $N_i$ and neuron $N_j$, and $E$ be the error function representing the root mean square error between desired output and computed output for each input training instance. Then the weight adaptation policy is formally given by

$$\delta W_{ij} = -\eta \frac{\partial E}{\partial W_{ij}}$$

where $\Delta W_{ij}$ denotes change in weight and $\eta$ is the learning rate. In Back-propagation algorithm, the weight adaptation for each layer is derived using the above equation.
Computation of $\Delta W_{ij}$ in the output layer of a multilayered feed-forward neural network is straightforward as the error function $E$ involves the weight $W_{ij}$. However, computation of $\Delta W_{ij}$ in the intermediate and input layers is not easy as the error function $E$ does not involve $W_{ij}$, and a chain formula of known partial derivatives is used to compute $\Delta W_{ij}$.

Once $\Delta W_{ij}$ computation is over, we add it to $W_{ij}$ to obtain its new value. The process of layerwise computation of weights always starts at the output layer, and continues up to the input layer, and this is usually referred to as one pass. Several passes are required for convergence of weights toward steady-state values. After convergence of the weights, the trained network can be used for the application phase. During this phase, the network is excited with a new instance.

One fundamental limitation of the back-propagation algorithm is trapping at local optima on the error (energy) surface. Several methods have been proposed to address the issue. The most common is adding momentum to the weight adaptation dynamics. This helps the dynamics to continue movement even after coming in close vicinity of any local optima. Once the dynamics pass the local optima, their speeds are increased so that the motion is continued until the global optimum is identified. Among the enormous work on emotion recognition using Back-propagation neural network, References 204–208 need special mention.

1.5.1.2 Radial Basis Function Based Neural Net  
Radial Basis Function (RBF) neurons employ a specialized basis function to map an input pattern $\vec{X}$ to two soft levels 0 and 1. A pattern classifier with $k$ classes usually has $k$ basis functions, designed to map an unknown input vector to one of $k$ classes. When a pattern falls in a class, its RBF function yields a value close to one. On the other hand when an input pattern does not fall in a class, the function returns a small value close to zero. Among the popular RBF functions, the Gaussian function is most common for its wide applications in science and engineering. Let $X_c$ be the center of an RBF function. Then for any input vector $X$, we define the RBF function $Y$ as given below:

$$Y = \exp(-\|X-X_c\|)$$

where $\|\|$ denotes an Euclidean norm.

A typical RBF neural net consists of two layers, the first layer being the RBF layer, and the last layer being realized by a perceptron neuron, the weights of which are determined by the perceptron learning algorithm. When an unknown input instance is supplied, the response of the first layered RBF neurons become close to zero for most of the neurons and close to one for one or fewer neurons. The weights generated by perceptron learning algorithm are later used to map an intermediate pattern into pattern class. A few research works employing Radial Basis Function Based Neural Network as a classifier for emotion recognition, include References 209–211.

1.5.1.3 Self-Organizing Feature Map Neural Net  
In a Self-Organizing Feature Map (SOFM) Neural Net, we need to map input patterns onto a 2D array of neurons
EMOTION CLASSIFICATION

based on the similarity of inputs with the patterns stored in individual neurons. The patterns stored by neurons have the same dimension as that of input patterns. These patterns are called weights of the respective neurons. A given input pattern is mapped onto a neuron with the shortest Euclidean distance. A neighborhood around the selected neuron is considered, and the weights of all the neurons in the neighborhood are adapted by the following equation:

\[ W_{ij}(t+1) \leftarrow W_{ij}(t) + \eta(X_k - W_{ij}(t)) \]

where \( W_{ij}(t) \) is the weight of neuron \((i,j)\) in the neighborhood of the selected neuron in the 2D array at time \( t \), \( X_k \) is the \( k \)th input vector, and \( \eta \) is the learning rate. After the weights are adapted, a new input vector is mapped onto the array by the distance criterion, and the process of neighborhood selection around the neuron and weight adaptation of the neurons in the neighborhood is repeated for all the inputs. The whole process of mapping input vectors onto a 2D array and weight adaptation of neurons is aimed at topological clustering of neurons, so that similar input vectors are mapped at close vicinity on the 2D array.

During the recognition phase, we need to retrieve one or more fields of a given vector, presuming that the remaining fields of the vector are known. Generally, the unknown vector has the same dimension to that of other input vectors used for weight adaptation. The unknown vector first is mapped onto a neuron in the 2D array based on the measure of minimum Euclidean distance between the input vector and all weight vectors, considering only the known fields of the unknown vector during Euclidean distance evaluation. The neuron having the best match, that is, having the smallest Euclidean distance between its weight vector and the unknown input vector is identified. The unknown fields of the input vector are retrieved from the corresponding fields of the weight vector of the selected neuron. SOFM can be used for emotion recognition from face [212], speech [213], EEG [214], as well as from gesture [215], for its high efficiency and accuracy as a classifier.

1.5.1.4 Support Vector Machine Classifiers Support Vector Machines have been successfully used for both linear and nonlinear classification. A linear SVM separates a set of data points into two classes with class labels \(+1\) and \(-1\). Let \( X = [x_1,x_2,...,x_n]^T \) be any point to be mapped into \(+1\), \(-1\) by a linear function \( f(X, W, b) \), where \( W = [w_1,w_2,...,w_n] \) is a weight vector and \( b \) is a bias term. Usually, \( f(X, W, b) = \text{Sign}(WX + b) = \text{Sign}(\sum w_j x_j + b) \). Figure 1.3 illustrates classification of 2D data points. In 2D, the straight line that segregates the two pattern classes is usually called a hyperplane. Further, the data points that are situated at the margins of the two boundaries of the linear classifier are called support vectors. Figure 1.3 describes a support vector for a linear SVM.

Let us now select two points \( X^+ \) and \( X^- \) as two support vectors such that for \( X = X^+ \), \( WX^+ + b = +1 \). Similarly, for \( X = X^- \), \( WX^- + b = -1 \). Now, the separation between the two support vectors lying in the class \(+1\) and class \(-1\), called marginal width is given by

\[ M = (WX^+ + b) - (WX^- + b)/\|W\| = 2/\|W\| \]
The main objective in a linear SVM is to maximize $M$, that is, to minimize $\|W\|$, which is same as minimizing $1/2W^TW$. Thus, the linear SVM can be mathematically described by

$$\text{Minimize } \phi(W) = \frac{1}{2}W^TW$$

subject to $y_i(WX_i + b) \geq 1$ for all $i$ (1.3)

where $y_i$ is either 1 or $-1$ depending on the class that $X_i$ belongs to.

Here, the objective is to solve $W$ and $b$ that satisfies the above equation. The solution to the optimization problem is not given here due to space limitation. However, inquisitive readers can have it in the standard literature. Linear SVM has a wide range of applications in supervised classifiers. It is currently one of the best popular algorithms for pattern classification. Many researchers choose SVM [18, 24, 26, 216–219] as a classifier in emotion classification for its higher accuracy.

1.5.1.5 Learning Vector Quantization  In scalar quantization, the random occurrence of a variable $x$ in a given range $[x_{\text{min}}, x_{\text{max}}]$ is quantized into few fixed levels. For example, suppose there are $m$ quantization levels, uniformly spaced in $[x_{\text{min}}, x_{\text{max}}]$. So, the quantization step height is determined as $q = (x_{\text{max}} - x_{\text{min}})/m$. The $k$th quantization level has a value $x_k = x_{\text{min}} + h.(k - 1)$. An analog signal $x$ having a value greater than $k$th quantization level, but less than $(k + 1)$th quantization level is quantized to the $k$th level. This particular feature of the quantization process is called truncation. Sometimes, we use roundoff characteristics of the quantizer. In case of roundoff, an analog signal having a value less than $x_k + q/2$ but greater than $x_k$ would be quantized to $x_k$. But if the analog signal $x > x_k + q/2$, but less than $x_{k+1}$, it will be quantized to $x_{k+1}$.

In vector quantization, vectors $V$ of dimension $n$ are quantized to fixed vectors $V_i$ of the same dimension. If all components of $V$ are close enough with respect to corresponding components of vector $V_i$, then $V$ would be quantized to $V_i$. Usually, such quantization is widely used in data compression. In learning vector quantization
(LVQ), we use a two-layered neural net. The first layer is used for data reception, while the second is a competitive layer, where one of several neurons only fires, and the weights are reinforced using the input. Let \( X_i \) be the \( i \)th input vector, whose components are mapped to the neurons at the input layer of the neural net. Let us assume that there are \( p \) neurons in the second layer. Now let the weight vector of the neurons in the second layer be \( W_1, W_2, \ldots, W_p \). Suppose

\[
\| W_k - X_i \| \leq \| W_j - X_i \| \text{ for all } j
\]

Then we would adapt \( W_k \) by the following update rule:

\[
W_k \leftarrow W_k + \eta (X_i - W_k)
\]

where \( \eta \) is the learning rate between \( (0, 1) \). The weights are thus reinforced by all the inputs. After the learning with all the input instances are over, we identify the unknown components of a given test instance vector by identifying the trained weight vector, with which the input instance has the smallest Euclidean distance, considering only the known components. The unknown fields of the selected weight vector are then used for subsequent applications. Researchers used LVQ as a classifier for emotion recognition from facial expression [220, 221] or speech [222].

1.5.2 Fuzzy Classifiers

Measurements obtained from facial expression, voice, and gesture/posture for emotion recognition often are found to be contaminated with various forms of uncertainty. For example, repeated measurements of facial, vocal, and bodily gestures of a subject experiencing the same emotion have wider variance. This is often referred to as an intrapersonal level of uncertainty. Further, when the measurements are taken from different subjects experiencing same/similar emotion, the variance in measurements is found to have a large value, causing interpersonal level of uncertainty. Classical type 1 fuzzy logic considers a single membership function to represent the uncertainty involved over a given measurement space, but fails to model the true spirit of intra- and interpersonal level of uncertainty [223]. Type 2 fuzzy sets provide an opportunity to represent both inter- and intrapersonal variations in uncertainty, and thus has immense scope in fuzzy classifiers, capable of correctly classifying emotions from measurements suffering from uncertainty.

In classical fuzzy rule–based classifiers, fuzzy rules are employed to map fuzzy encoded measurements into emotion classes with different degrees of certainty. Thus individual emotions support a given set of emotional features to different degrees, and naturally the class with the highest support is considered as the winner. Type 2 fuzzy rules on the other hand map a set of imprecise fuzzy encoded measurements obtained from different sources into emotion classes. The class offering maximum support to the measurement space is considered as the winning class.

An alternative approach to fuzzy pattern recognition is to cluster a set of features based on their similarity. Since measurements are noisy, data points lying on the
boundary of two classes can be categorized to both the classes with certain degrees. Fuzzy C-means clustering algorithm has widely been used as a basic tool of pattern clustering to overcome clustering of data points suffering from noisy measurements or incomplete specification of data dimensions. A few important research works employing fuzzy logic in classifiers include [60,224,225].

1.5.3 Hidden Markov Model Based Classifiers

Let $X$, $Y$, and $Z$ be three random variables where they may take up any value from $x_1, x_2, \ldots, x_n$, $y_1, y_2, \ldots, y_m$, and $z_1, z_2, \ldots, z_r$, respectively. Suppose $Y$ depends on $X$ and $Z$ depends on $Y$. So, we can represent the dependence relationship among $X$, $Y$, and $Z$ by a directed graph, where $X$ is a predecessor of $Y$ and $Y$ is a predecessor of $Z$. Now suppose, by experiments we have the conditional probabilities: $P(Y/X)$ and $P(Z/Y)$ for any $X$, $Y$, and $Z$. Now in a Markov process we consider the state-transition probabilities in one level only, that is, we consider $P(Z/Y, X) = P(Z/Y)$.

In an HMM, the probability of occurrence of a class is determined by a sequence of state transitions. For example, suppose if $X = x_2$ then $Y = y_3$, and if $Y = y_3$ then $Z = z_4$. Suppose, the sequence of state transitions from $X = x_2$ through $Y = y_3$ and $Z = z_4$ denotes class 1 of a pattern recognition problem. This state transition probability is $P(Z = z_4/Y = y_3) \cdot P(Y = y_3/X = x_2)$. Now, if $P(X = x_2)$ is the probability of occurrence of $X = x_2$ through $Y = y_3$ is given by $P(X = x_2) \cdot P(Y = y_3/X = x_2) \cdot P(Z = z_4/Y = y_3)$. Thus for all known sequences passing through $X$, $Y$, and $Z$, we know the probability of the pattern class.

Now, for an unknown sequence, we compute the state transition probability of the sequence, and check whether the probability of the sequence matches enough with the probability of the standard nearest sequence. If so, we consider the unknown sequence to lie in the class of the nearest matched sequence. HMM is vastly used in research for recognition of emotion from face [25], speech [226–228], and sometimes from both face and speech [229–231].

1.5.4 k-Nearest Neighbor Algorithm

k-Nearest Neighbor algorithm is a simple method to determine the class of an unknown pattern. It presumes that the given data points are pre-classified and the label of each class is known. Now, for an unknown pattern, represented by a data point $X$, the distance between $X$ and all nearest neighbors $Y_j$ of $X$ are identified, and the count of $Y_j$ lying in different classes is determined. The class having the largest number of nearest neighbors of $X$ is declared as the class of the unknown data point $X$.

The distance metric selection is an important problem in the k-NN algorithm. Usually, Euclidean distance is used in most of the literature [232]. However, when the data points have a large dimension or the components are not scaled properly, the distance between any two data points sometimes is too large, and consequently the results of classification is not free from errors. The given set of data points therefore should be normalized by scaling $d_{ij}$ by $d_i/	ext{Max}_j(d_{ij})$, where $d_{ij}$ is the $j$th component of the $i$th data point $d_i$. 

k-NN algorithm works fine for lower dimensional data, typically when dimension is less than 5. It has popularity in the pattern recognition community particularly for its simplicity. The complexity of the algorithm grows with an increase in the dimension of the data points. To avoid this increase in computational complexity, feature reduction algorithm is first employed to identify the independent components of the data points by using a feature reduction algorithm, and k-NN is applied on the data classes developed with reduced dimensionality. k-NN has been used in many research works as a classifier to recognize emotion [79,90,233].

1.5.5 Naïve Bayes Classifier

Suppose, there are \( n \) objects, each having a set of \( m \) features \( f_1, f_2, \ldots, f_m \) and the objects are classified into \( k \) classes: \( c_1, c_2, \ldots, c_k \) based on the similarity of their features/attributes. Now, by Bayes’ conditional probability theory, we have the conditional probability

\[
P(c_j \mid f_1, f_2, \ldots, f_m) = \frac{P(f_1, f_2, \ldots, f_m \mid c_j)P(c_j)}{P(f_1, f_2, \ldots, f_m)}
\]

where the probabilities have their standard meaning. Now, for any \( j \), the expression on the right-hand side contains a constant denominator: \( P(f_1, f_2, \ldots, f_m) \). So, we instead of computing \( P(f_1, f_2, \ldots, f_m) \), simply represent it by a scalar constant \( 1/A \). Therefore, the last result may be reconstructed as

\[
P(c_j \mid f_1, f_2, \ldots, f_m) = A \cdot P(f_1, f_2, \ldots, f_m \mid c_j)P(c_j)
\] (1.4)

Now, \( P(f_1, f_2, \ldots, f_m \mid c_j) \) denotes the probability of the joint occurrence of \( f_1, f_2, \ldots, f_m \), given the object class \( c_j \). However, computation of \( P(f_1, f_2, \ldots, f_m \mid c_j) \) is not simple. One simple assumption is to consider the features independent, that is,

\[
P(f_1, f_2, \ldots, f_m \mid c_j) = P(f_1/ \mid c_j)P(f_2/ \mid c_j)\ldots P(f_m/ \mid c_j) = \prod_{i=1}^{m} P(f_i/ \mid c_j)
\] (1.5)

Combining (1.3) and (1.4) we obtain,

\[
P(c_j \mid f_1, f_2, \ldots, f_m) = A \cdot P(f_1, f_2, \ldots, f_m \mid c_j)P(c_j) = A \cdot \prod_{i=1}^{m} P(f_i/ \mid c_j)P(c_i)
\]

Now, from experimental instances, if we can prepare a table \( P(f_i/ \mid c_j) \) for all \( i,j \), we can determine \( P(c_j \mid f_1, f_2, \ldots, f_m) \) for all \( j \) and identify the class \( k \), for which

\[
P(c_k \mid f_1, f_2, \ldots, f_m) \geq P(c_j \mid f_1, f_2, \ldots, f_m) \text{ for all } j
\]

The class \( k \) is the desired class for the given combination of features. This is how a Naïve Bayes classifier works. The current literature employing Naïve Bayes Classifier to recognize emotion includes References 234–236.
1.6 MULTIMODAL EMOTION RECOGNITION

Emotion of a subject is recognized from his or her facial expression, voice, gesture, and posture. The term modality refers to modes of recognition. Early research in emotion recognition stressed the importance of single modality. Single modality, however, cannot always provide sufficient information for emotion recognition. Current research on emotion recognition thus is targeted to multimodal information extraction to improve the reliability of the recognition process.

One common metric to represent the performance of the recognition process is to construct a confusion matrix, where each grid \((i, j)\) of this matrix represents the percentage of emotional expression correctly classified in class \(j\), when the actual emotion is \(i\). Experiments undertaken on different modality of the same subjects reveal that the confusion matrices constructed for each modality have significant differences, indicating that the individual modality of classification is not highly reliable.

Multimodal fusion is to integrate all single modalities into a combined single representation. There are basically two types of fusion techniques that have been used in most of the literature to improve reliability in emotion recognition from multimodal information. They are (1) feature-level fusion and (2) decision-level fusion. Feature-level (early) fusion classifies emotion of subjects from their individual modalities in the first step, and then attempts to fuse the results of unimodal classification with an aim to reduce uncertainty in classification. One simple method of fusion is to check whether the individual classifiers for multiple modalities yield the same class for a subject. If so, the decision about the emotion class of the subject is certain. However, if the classifiers produce conflicting class labels for a subject, then the class label obtained from majority of the classifiers is honored. Decision-level (late) fusion feeds all measurements obtained from sensors employed for different modalities to a pretrained classifier, which in turn determines the class label of the unknown measurement.

In Reference 173, the authors applied both feature-level and decision-level fusion on face and bodily gestures and concluded that feature-level fusion achieves better performance than decision-level fusion. On the other hand, in Reference 172, the authors observed that the overall performance of the feature-level bimodal classifier and decision-level bimodal classifier is similar. But by analyzing with confusion matrix, they also observed that the recognition rate for each emotion type was totally different. In the decision-level bimodal classifier, the recognition rate of each emotion is increased (except happiness, which is decreased by 2%) compared to the facial expression classifier. Recognition rate of emotion anger and neutral are significantly increased in the feature-level bimodal classifier whereas recognition rate of happiness is decreased by 9% by using the same. Therefore, they concluded that “the best approach to fuse the modalities will depend on the application.”

1.7 STIMULUS GENERATION FOR EMOTION AROUSAL

Elicitation of emotion usually is performed by presenting selective stimuli to the subjects. Selection of the modality of emotion elicitation and identification of the
right stimuli for a given modality are equally important. Among the well-known modalities of emotion elicitation, visual, audio, audiovisual, music are most popular. Traces of using ultrasonic signals for emotion elicitation are also found in the literature [237]. Proximity signals, such as touch of a mother to stop crying of babies, have also been used in the literature for emotion control.

Considering a particular modality of emotion elicitation, we can identify some fundamental basis for selection of stimulus. In Reference 60, the authors sought the opinion of several people to determine the candidate audiovisual stimulus, which seems to be most effective for elicitation of a particular emotion. To briefly outline their scheme, suppose we have \(n\) audiovisual stimuli presented before a group of \(m\) subjects. Each subject is asked to provide his or her feeling about possible percentage of elicitation of five emotions by the stimuli. After the score sheet containing the opinions of \(m\) subjects for five emotions excited by \(n\) stimuli is prepared, the sum of percentage support obtained from \(m\) subjects for each stimulus favoring individual emotion class is determined. The emotion class that receives highest support for each stimulus is selected for elicitation of that emotion.

Selection of stimuli and their modalities are important issues for arousal of emotion. The modalities include stimulation by audio, visual, audiovisual, consolation message, music, and dance. Among these modalities, significant research has been undertaken on audio [238], visual [239, 240], audiovisual [241–243] and music [244–248] based stimuli for emotion arousal. In Reference 244 the authors studied the effect of certain features in music as a qualitative measure of emotion arousal. They noticed that pitch range, tone quality, rhythmic regularity, and tempo are some of the important features capable of originating and regulating emotion. Some well-known music pieces are often used for generating particular emotion in subjects. Baumgartner et al. [240] selected some classical orchestral pieces like Gustav Holst: Mars—The Bringer of War from the Planets, Samuel Barber: Adagio for Strings, and Beethoven: Symphony No. 6 (3rd mvt.) for arousal of three basic emotions: fear, happiness, and sadness. The tones of Holst evoke fear, music of Barber evokes sadness, and the one by Beethoven evokes happiness.

Visual mode of excitation includes voiceless movie, particularly action movies describing thematic issues of a particular real-life context. The modality of video presentation, context, and climax keep the subject attracted toward the stimulus, and the context of the movie excites a specific faculty of the mind, arousing an emotion. For example, a set of consecutive pathetic scenes causes an arousal of sadness, while a scene of amusements causes relaxation/happiness. In Reference 240, the authors have shown 48 still pictures containing humans or human faces to 70 subjects for arousal of three basic emotions: fear, happiness, and sadness. They have selected pictures of a man attacking a woman with a knife or a man pointing a pistol to the viewer for arousing fear. Pictures of a man holding his smiling baby, laughing children playing on the beach, or athletes in a victory pose are used for making subjects happy. For arousing sadness, picture of a crying little boy standing in front of a destroyed house or a couple standing at a gravestone are shown.

Audiovisual modality of excitation has been found to very effective for control of emotional state-transitions. The onset time (i.e., the duration of stimulation after
which the brain responds) and offset time (i.e., the minimum time brain remains excited after removal of the stimulus) \cite{31, 249} are important issues in audiovisual stimulation. These times are not always constant for a given subject. They are sensitive to both context and climax.

### 1.8 VALIDATION TECHNIQUES

After a system design is complete, we go for validation of the system performance using test samples. In case performance of the system is found satisfactory, the system design is certified for field applications. Otherwise, we should identify the free system variables/parameters, which are not direct inputs or outputs, for tuning. Machine learning and/or evolutionary algorithms are often used to tune the parameters of physical systems in order to improve the system performance with known test samples. In Reference 60, the authors have employed evolutionary algorithm to adapt the parameters of fuzzy membership functions to improve the classification accuracy of the fuzzy rule-based classifier. The classifier in that paper has been tested with facial features of subjects with known emotion classes.

A good system validation requires identifying right test instances taken from all classes of emotion with wider variance in feature space even for the same class. If the test instances are misclassified, we need to adapt system parameters. Selection of non-input/output parameters having a wider range of variation with a power of controlling the performance of the classification algorithm is also an important issue. Lastly, the evolutionary/learning algorithm to be incorporated should have the ability to tune system parameters with a motive of improving the performance index of the classifier. Usually system validation is done offline. So, no restriction is imposed on the convergence time of the validation algorithm. However, a fast algorithm is always preferred without allowing degradation in system performance.

In Reference 177, authors used leave-one-out cross-validation method for selecting the correct classifier between SFFS, FP, and a combination of SFFS and FP, called SFFS-FP. In each case, the feature vector per emotion to be classified was excluded from the dataset before the SFFS, FP, or SFFS-FP was run. The best set of features or best transform (determined from the training set) was then applied to the test data point to determine classification accuracy. They observed that SFFS-FP improves the performance and provides the highest recognition rate on eight classes of emotion.

Busso et al. \cite{78}, compares the performance of their proposed approach with the conventional classifier when there is a mismatch between the training and testing conditions. For this purpose, they separate the emotional databases into two language groups: English and German. One of these groups is used for training, and the other one for testing. The system is then trained and tested with different databases recorded in three different languages (English, German, and Spanish) to validate the robustness of the proposed approach. It has been seen that the performance of the conventional classifier without the neutral models decreases up to 17.9%. So they concluded that “using neutral models for emotional recognition is hypothesized to be more robust and, therefore, to generalize better than using a direct emotion classification approach.”
1.8.1 Performance Metrics for Emotion Classification

Performance metrics are used to measure the performance of an emotion classifier. The simplest metric for performance measure is classification accuracy, which indicates the percentage of success in emotion classification from a given set of measurements of the subjects during the period of their emotion arousal. Classification accuracy cannot provide the subjective details about how many instances of emotional expressions falling under an emotion class are misclassified in all other emotion classes. For example, if out of 60 facial expressions, 30 appear in anger, 20 in fear, and 10 in anxiety classes, then we understand that the set of features is either incomplete or the emotional expressions are not sufficiently distinguishable. Confusion matrix offers the details of percentage misclassification of an emotion into other classes in percentage or absolute values, and so it is a popularly used metric to represent classification accuracy in emotion recognition problem.

Statistical tests are undertaken to compare the relative performance of emotion classifiers. Two types of tests are generally performed. The first type compares the performance of two classifiers, while the latter type determines ranks of individual classifiers based on some metric, usually classification accuracy. The merit of using these statistical tests is to ensure robustness in performance analysis. The tests require some measure of individual performance of a classifier tested with different databases, and ultimately determine the relative performance of the classifiers as obtained from all the databases together. Among the well-known statistical tests, McNemar’s test and Friedman test [250, 251] need special mention. McNemar’s test considers two classifier algorithms at a time, and identifies the relatively better between them based on a specific measure determined by the count of misclassified items by the individual algorithms. Friedman test considers all the classifier algorithms together and ranks them best on the individual performance of the algorithms.

1.9 SUMMARY

Pattern recognition at present is an established discipline of knowledge. A general pattern recognition problem comprises three or more steps: feature extraction, feature reduction, and classification. Emotion recognition falls under the basic category of pattern recognition problems. Thus, the basic steps of pattern recognition are equally applicable here. However, the main challenge in emotion recognition lies in extraction of essential features, and selecting a suitable classifier. The chapter provides a thorough review of the existing techniques of feature extraction and classification, commonly used in emotion recognition.

A review of the existing techniques of emotion recognition from a pattern classification perspective is given in the chapter. Special emphasis is given to modality-based recognition as for individual modalities the feature extraction and classification techniques have significant differences. An outline to different validation techniques for emotion classification is given here. In addition, the statistical techniques used to compare the relative performance in emotion classification are narrated at the end of the chapter.
INTRODUCTION TO EMOTION RECOGNITION

On occasion single modality cannot correctly yield good accuracy. Multimodality becomes an important necessity under the context. The chapter provides an extensive discussion on multimodal emotion recognition. Stimulus generation for elicitation of emotion is an open problem, and very few works in this regard are available in the literature. The chapter provides a thorough overview on stimulus generation. Lastly, the chapter cites an exhaustive list of references [1–264] to emotion research, including both fundamental research and applications.

REFERENCES

REFERENCES


INTRODUCTION TO EMOTION RECOGNITION


32 INTRODUCTION TO EMOTION RECOGNITION


REFERENCES


INTRODUCTION TO EMOTION RECOGNITION


REFERENCES


INTRODUCTION TO EMOTION RECOGNITION


REFERENCES


INTRODUCTION TO EMOTION RECOGNITION


REFERENCES


INTRODUCTION TO EMOTION RECOGNITION


REFERENCES


INTRODUCTION TO EMOTION RECOGNITION


[242] M. Ghosh, A. Chakraborty, A. Konar, and A. Nagar, “Detection of chaos and limit cycles in the emotional dynamics from the facial expression of subjects stimulated with audio-visual movies,” in Computational Intelligence, Control and Computer Vision in Robotics and Automation (ed. B. Subudhi), Narosa Publisher.

REFERENCES


44 INTRODUCTION TO EMOTION RECOGNITION


AUTHOR BIOGRAPHIES

Amit Konar received the B.E. degree from Bengal Engineering and Science University, Howrah, India, in 1983, and the M.E., M.Phil., and Ph.D. (Engineering) degrees from Jadavpur University, Kolkata, India, in 1985, 1988, and 1994, respectively. In 2006, he was a visiting professor with the University of Missouri, St. Louis. He is currently a professor with the Department of Electronics and Telecommunication Engineering (ETCE), Jadavpur University, where he is the founding coordinator of the M.Tech. program in Intelligent Automation and Robotics. He has supervised fifteen Ph.D. theses. He has over 200 publications in international journals and conference proceedings. He is the author of eight books, including two popular texts Artificial Intelligence and Soft Computing (CRC Press, 2000) and Computational Intelligence: Principles, Techniques and Applications (Springer, 2005). He serves as the Associate Editor of IEEE Transactions on Systems, Man and Cybernetics, Part A and IEEE Transactions on Fuzzy Systems. His research areas include the study of computational intelligence algorithms and their applications to the domains of electrical engineering and computer science. Specifically, he has worked on fuzzy sets and logic, neurocomputing, evolutionary algorithms, Dempster–Shafer theory, and Kalman filtering, and has applied the principles of computational intelligence in image understanding, VLSI design, mobile robotics, pattern recognition, brain–computer interfacing, and computational biology. He was the recipient of All India Council for Technical Education (AICTE) accredited 1997–2000 Career Award for Young Teachers for his significant contribution in teaching and research.

Anisha Halder received the B.Tech. degree in Electronics and Communication Engineering from the Haldia Institute of Technology, Midnapore, India, and the M.E. degree in Control Engineering from Electronics and Telecommunication Engineering Department, Jadavpur University, Kolkata, India, in 2007 and 2009 respectively. She is currently pursuing her Ph.D. degree under the guidance of Prof. Amit Konar and Dr. Aruna Chakraborty at Jadavpur University.
Her principal research interests include artificial intelligence, pattern recognition, cognitive science, and human–computer interaction. She is the author of over 25 papers published in top international journals and conference proceedings.

Aruna Chakraborty received the M.A. degree in Cognitive Science and the Ph.D. degree in Emotional Intelligence and Human–Computer Interactions from Jadavpur University, Kolkata, India, in 2000 and 2005, respectively. She is currently an associate professor in the Department of Computer Science & Engineering, St. Thomas’ College of Engineering & Technology, Kolkata. She is also a visiting faculty of Jadavpur University, where she offers graduate-level courses on intelligent automation and robotics, and cognitive science.

She has coauthored a book with Prof. A. Konar titled Emotional Intelligence: A Cybernetic Approach (Springer, Heidelberg, 2009). She serves as an editor to the International Journal of Artificial Intelligence and Soft Computing, Inderscience, UK. Her current research interests include artificial intelligence, emotion modeling, and their applications in next-generation human–machine interactive systems. She is a nature lover, and loves music and painting.