CHAPTER 1
INTRODUCTION AND SCOPE OF THE WORK

1. INTRODUCTION

Humans interrelate with each other mainly through speech, also through body gestures, to highlight a certain part of speech and display of emotions. In an effort to understand and categorize emotions, psychologists and engineers have tried to analyze facial expressions, vocal emotions, gestures, and physiological signals.

From the video images acquired from built-in cameras, and from speech waveforms collected from on-board microphones, this information can be used to teach computers to recognize human emotions. A normal two-way interaction between the human and the computer through multiple modalities is represented in Figure 1.1. In this diagram, one of the inputs to the computer is a video, from which gaze, posture, gestures, and facial and lip movements can be extracted. Computers may learn to recognize gestures, postures, facial expressions, eye contact. Similarly, speech and voice (audio) through the microphone may convey linguistic as well as paralinguistic information. On the output side, the computer may appear in the form of an “agent”- a computer-animated face. This “agent” can speak to the human
through a created speech and display corresponding facial and mouth movements on the screen.

Fig.1.1 Multimodal human-computer interaction
Even if they are not openly presented in the Figure 1.1, some other modalities such as tactile or physiological signals can also be used in concurrence with the video and audio signals. The term “Emotional expression” means any outward expression that arises as a response to some stimulus event. These may include typical expressions such as a “smile” to show that one is happy, or to show one likes what one sees.

1.1 LABELLING EMOTIONS

For labelling the emotions in discrete categories, human must choose from a prearranged list of word labels, such as joy, fear, love, surprise, sadness. A problem with this approach is that the stimuli may hold blended emotions. The choice of words may be too restrictive, or culturally dependent.

Emotions can also be defined in multiple dimensions or scales. Instead of picking discrete labels, viewers can specify their impression of each stimulus on several nonstop scales, pleasant-unpleasant, attention-rejection, simple-complicated. The 2 common scales are valence and arousal. Valence describes the niceness of the stimuli, with positive/pleasant on one end, and negative/unpleasant on the other end. Happiness has a positive valence. Disgust has a negative valence. The extra dimension is arousal or activation. Sadness has low arousal. Surprise has high arousal level.
Lang, 1995, described the emotion labels in different form could be plotted at several positions on a 2D plane, covered by the two axes, to construct a 2D emotion model. Scholsberg, 1954, suggested a 3-D model in which the author had attention and rejection.

Most researchers use pattern recognition approaches for recognizing emotions, by labeling emotions into different states. They used different modalities as inputs to the emotion recognition models.

1.2 FACIAL EXPRESSION RECOGNITION STUDIES

During 1970s, Paul Ekman and his colleagues performed wide studies of human facial expressions (Ekman 1994). They found proof to support universality in facial expressions. These “universal facial expressions” are happiness, sadness, anger, fear, surprise, and disgust. They studied facial expressions in various cultures, with preliterate cultures, and found much unity in the expression and recognition of emotions on the face. They observed difference in expressions as well, and proposed that facial expressions are governed by “display rules” in different social contexts.

Matsumoto, 1998, reported the discovery of a seventh universal facial expression, contempt. Izard, 1994, explained that the babies displayed a wide range
of facial expressions without being taught, thus suggesting that these expressions are distinctive.

Ekman and Friesen (Ekman 1978) described a set of action units (AUs). They developed Facial Action Coding System (FACS) to code facial expressions, the movements on the face. Each AU has some related muscular basis. Each facial expression is described with a combination of AUs. By following set of prescribed rules, this system of coding facial expressions is done manually. Frequently at the peak of the expression, the inputs are still images of facial expressions.


Mase, 1991, used optical flow (OF) to recognize facial expressions. They used image processing techniques to recognize facial expressions.

Lanitis, 1995, used a flexible shape and appearance model for image coding, person classification, pose recovery, gender recognition and facial expression recognition.
Black, 1995, used local parameterized models of image motion to recover non-rigid motion. Once recovered, these parameters are fed to a rule-based classifier to recognize the six basic facial expressions.

Yacoob, 1996, computed optical flow and used similar rules to classify the six facial expressions. Rosenblum, 2000, computed optical flow of regions on the face, and then applied a radial basis function network to classify expressions.

Essa, 1997, used an optical flow region-based method to recognize expressions.

Otsuka, 1997, computed optical flow at the first time, then computed 2D Fourier transform coefficients, which were used as feature vectors for a hidden Markov model (HMM) to classify expressions. The trained system was able to recognized one of the six expressions. They used the tracked motions to control the facial expression of an animated Kabuki system.


Martinez, 1999, presented an indexing approach based on the identification of frontal face images under various illumination conditions, facial expressions, and occlusions. A Bayesian approach was used to find the best match between the local
observations and the learnt local features model. HMM was used to achieve good recognition.

Oliver, 2000, used lower face tracking to extract mouth shape features. He used the features as inputs to an HMM based facial expression recognition system. He identified the expressions happy, sad, and neutral and an open mouth.

Chen, 2000, used a suite of static classifier to recognize facial expressions, reporting on both person-dependent and person-independent results.

Cohen, 2003, described classification schemes for facial expression recognition in two types of settings, dynamic and static classification. Based on the tracking results of that frame, the static classifier classifies a frame in a video to one of the facial expression categories. For the dynamic setting, Cohen 2004 used a multi-level HMM classifier that combines the temporal information, and allows not only to achieve the classification of a video segment to the corresponding facial expression, but also to automatically segment an arbitrary long sequence to the different expression. These methods are similar in the general sense that they extracted some features from the images and fed features into a classification system, and the outcome is one of the preselected emotion categories.

The video processing falls into two broad categories. The first is “feature-based,” where one tries to detect and track specific features such as the corners of the mouth, eyebrows. The other approach is “region-based” in which facial
motions are measured in certain regions on the face such as the eye/eyebrow and mouth regions. People have used different classification algorithms to categorize these emotions.

Bassili, 1979, compared several facial expression recognition algorithms. They stated that these algorithms perform well, compared to trained human recognition of about 87%. Ueki, 1995, extracted AUs and used neural networks (NN) to analyze the emotions, mapping seventeen AUs to two dimensions using an identity mapping network, and this showed resemblance of the 2D psychological emotion models. Morishima, 1995, used a 3D emotion model in order to deal with transitions between emotions, and claimed correlation to the 3D psychological emotion model (Schlosberg 1954).

As reported by Ekman (1978), anger and disgust are commonly confused in judgment studies. Fear and surprise are commonly confused. Because of sharing the similar facial actions these confusions occurs. Surprise is occasionally mistaken for interest. In the computer recognition studies, some of these confusions are observed (Black 1995, Yacoob 1996, Cohen 2003).

1.3 VOCAL EMOTION RECOGNITION STUDIES

The vocal aspect of a communicative message carries various kinds of information. Studies in emotional contents in speech (Murray 1993, Chiu 1994,

Williams, 1972, studied the spectrograms of real emotional speech and compared them with acted speech. They found similarities which suggest the use of acted data. Murray, 1993, reviewed the findings on human vocal emotions. They also constructed a synthesis-by-rule system to incorporate emotions in synthetic speech (Murray 1996). Chiu, 1994, extracted five features from speech and used a multilayered neural network for the classification. For 20 test sentences, they were able to correctly label all three categories. Dellaert, 1996, used 17 features and compared different classification algorithms and feature selection methods.

Petrushin, 1998, compared human and machine recognition of emotions in speech and achieved similar rates for both (around 65%). Scherer, 1996, performed a large-scale study using 14 professional actors. In their study, they extracted 29 features from the speech. According to Scherer, human ability to recognize emotions from purely vocal stimuli is about 60%. They pointed out that sadness and anger is best recognized, followed by fear and joy. Disgust is the worst. Chen, 2000, implemented a rule-based method for classification of input audio data into one of the following emotion categories: happiness, sadness, fear, anger, surprise, and dislike. The input data contained 2 speakers; speaker one speaks Spanish and
the other one Sinhala. The choice of these languages was such that the subjective judgments were not influenced by the linguistic content as the observers did not comprehend either language. Each speaker was asked to speak 6 different sentences for each emotion, and the contents of the sentences were related in most of the cases to one category and some of them could be applied to two different categories. From the audio signals pitch, intensity, and pitch contours were estimated as acoustic features, using predefined rules. Many studies used the “Ekman six” basic emotions. The reasons for using these basic six categories are often not justified. Murray et al., 1993, describes mostly qualitative characteristics associated with these emotions and listed in relation to the neutral voice.

1.4 EMOTION RECOGNITION FROM PHYSIOLOGICAL SIGNALS

Emotion consists of more than outward physical expression; it also consists of internal feelings and thoughts, as well as other internal processes of which the person having the emotion may not be aware. Still, these physiological processes can be naturally recognized by people. A stranger shaking hand can feel its clamminess (related to skin conductivity).

Physiological pattern recognition of emotion has important applications in medicine, entertainment, and human-computer interaction (Picard 2001).
Physiological pattern recognition can potentially help in assessing and quantifying stress, anger, and other emotions that influenced the health.

One of the big questions in emotion theory is whether distinct physiological patterns accompany each emotion. The physiological muscle movements comprising what looks to an outsider to be a facial expression may not always correspond to a real underlying emotional state. This relation between the bodily feelings and externally observable expression is still an open research area, with a history of controversy.

1.5 MULTIMODAL APPROACH TO EMOTION RECOGNITION

The research in facial expression recognition and vocal affected recognition has been done basically independent of each other. The aforementioned works in facial expression recognition used static photographs or video sequences where the subject exhibits only facial expression without speaking one word. The works on vocal emotion detection used only the audio information. There are situations where people speak and display facial expressions at the same time. A person said hello with a smile. Pure facial expression recognizers may fail because the mouth movements may not fit to the description of a pure “smile.”

The success of the research efforts has shown that fusing audio and video for detection of discrete events using probabilistic graphical models is possible.
Therefore, while the network shown is static, it can be extended to be a dynamic Bayesian network in a straightforward manner. In a probabilistic manner, the network topology combines the two modalities. The top node is the class variable for recognized emotional expression. It is affected by recognized facial expressions, recognized vocal expressions, recognized keywords that have an affective meaning, and by the context in which the system operates. From the person’s audio track, audio features are extracted. Then Vocal emotions are recognized. Using video, by facial features tracking method, the facial expressions are recognized, but the recognition is also affected by a variable that specifies, whether the person is speaking or not. Recognizing a person in speaking uses both visual cues, audio features (Garg 2003). The parameters of the implemented network can be learnt from data, or manually set for some variables. Inferring the human emotional expression can be achieved even when some pieces of information are missing, like, noisy audio, or loses the face while face tracking.

Another issue which makes the problem of emotional expression recognition even more difficult to solve in a general case is the dependency of a person’s behavior on his/her personality, cultural, and social vicinity, current mood, and the context in which the observed behavioral cues were encountered. One source of help for these problems is machine learning, rather than using a priori rules to interpret human behavior. One can potentially learn application, user, and context
dependent rules by watching the user’s behavior in the sensed context. This leads to another advantage of probabilistic graphical models. Well-known algorithms exist to adapt the models, and it is possible to use prior knowledge when learning new models. A prior model of emotional expression recognition trained based on a certain user. It can be used as an initial point for learning a model for another user, or for the same user in a different context. Though context sensing and the time needed to learn appropriate rules are significant problems in their own right, many benefits could come from such an adaptive affected-sensitive HCI tool.

1.6 IMPLEMENTED ARCHITECTURE FOR FACIAL EMOTION EXPRESSION CLASSIFICATION

Working Details of implemented Architecture (Figure 1.2):

Step 1: Frames are extracted from a video.

Step 2: Important features are extracted from successive frames that belong to one second.

Step 3: The features are trained using the implemented RBF/ESNN/CPN/BPA and final weights are stored in a database.

Step 4: In the testing process, step 1 and step 2 are adopted. The extracted features are processed with final weights of the RBF/ESNN/CPN/BPA, to get an output in the output layer of the RBF/ESNN/CPN/BPA.
**Step 5:** The output is compared with a threshold value, to decide the category to which the particular emotion the facial expression belongs.

![Proposed Architecture for Facial Emotion Expression Classification](image-url)

*Fig. 1.2 Proposed Architecture for Facial Emotion Expression Classification*
1.7 PROBLEM STATEMENT

Automated facial expression analysis is a vast research field. Pantic, 2000, provided a comprehensive overview of the state of the art and define problems to expression analysis.

1. Face detection in an image or image sequence.
2. Facial expression data extraction.
3. Facial expression classification. Face detection is dealt with implicitly. For still images, it is assumed that these are faces. For live video, face tracker automatically localizes facial features only if a face is present in the video stream.
4. The pose of the subject, the position of the face can cause some parts of the face such as eyes and nose occluded in the captured image (Yang, 2002).
5. The system should be invariant to different lightening conditions and distraction as glasses, change in hair style, facial hair, moustache, and beard. The features like beard, moustache, and glasses may not result in the correct facial expression because these features are subjects to change.
6. Person’s affective natural changes with time could cause some difficulties in the machine’s correct classification
7. Faces may be partially occluded by other objects such as scarf and hat.
1.8 IMPLEMENTED METHODOLOGIES

In this research, a systematic approach has been developed to train ANN topology with different algorithms for emotion classification. The algorithms are as follows:

1. Radial basis function Network (RBF).
2. Echo state neural network (ESNN).
3. Counter propagation neural network (CPN).

1.9 NEED FOR THE RESEARCH

1. To minimize difficulties in facial emotion expression (FEE) recognition due to the variation of facial expression across the human population.
2. To analyze facial expression feature extraction methods.
3. To compare the performance of facial expression classification scheme.

1.10 MOTIVATION FOR THE RESEARCH

1. For human, analysis of facial expression is simple but for computers it is a sophisticated problem that requires complex algorithms and techniques in image analysis and high dimensional pattern recognition.
2. The development of a universal anatomical model for faces across different cultures, age groups and demographical origins is also a different task.

3. This work addresses the problem of detecting the facial expressions by analyzing the video images.

4. The work is motivated to reduce the intensive computations in the previous work to achieve better accuracy and results.

1.11 SCOPE OF THE WORK

The emotion classification through facial expression is carried from the frames collected in a video. The video is taken with the persons expressing different expressions.

1.12 OBJECTIVES OF THE WORK

1. To recognize the different types of expressions of a face.

2. To implement artificial neural networks (ANN) for classification of emotions from the facial expressions.
1.13 ORGANIZATION OF THE THESIS

Chapter 2 presents detailed literature review on emotion classification.

Chapter 3 presents details of facial emotions expressions data collections, facial tracking and feature extraction.

Chapter 4 presents the implementation of proposed BPA / RBF/ ESNN / CPN algorithms for emotion classification.

Chapter 5 presents results and discussion in classifying the correct emotions.

Chapter 6 presents the conclusions and future scope of the work.