1. INTRODUCTION

1.1 INTRODUCTION

In recent years there has been a growing interest for achieving an effective human-computer interaction (HCI). A computerized information booth can answer to an user’s request using virtual talking heads. Whenever the computer is talking to an user, we prefer emotional faces instead of mask-like faces. If the user can select the person who is answering his question, it will be better. Virtual talking characters make more fun [8], [9]. There are many ways humans can express their emotions. Emotions can be expressed by means of speech, facial expressions, body movement etc. Facial expressions are the most powerful, natural and immediate means to communicate emotions and intentions [7], [11].

Facial Image Synthesis finds major applications in animations for Entertainment, Education, Training for hearing impaired, Broadcasting, Commerce, Navigational Aids, Chatting etc [4], [5], [10], [50]. Another interesting application is in robotics, where a robot’s face can generate emotions intelligently.

Facial animation has very much importance, eg. a nonhuman cartoon character can also be animated for speaking with emotions. In animation films, the facial expressions are highly exaggerated. The faces can express emotions before they verbalize. Facial images are difficult to interpret because they are highly variable with respect to individual appearance, facial expressions, lighting etc. When a person speaks, his emotions vary from sentence to sentence or even from words to words. When a person speaks without any emotions, the only facial part which undergoes movement is region around lip. When he is speaking with emotions, he moves all the facial regions such as lips, cheeks, eyes, eyebrows etc. We have to consider all these factors. Facial expressions are controlled by individual movements of facial muscles [2], [8].

In a text to Facial Image Synthesis (FIS) scheme, the lip movement depends on the language that is used. Indian languages are phonetic. "what we speak is what we
write”, unlike some foreign languages like English. So pronunciation can be broken to letters rather than words. The basic unit of acoustic speech is called a *phoneme* [44]. A viseme is the smallest visible unit of speech. There are many acoustic sounds that are visually ambiguous. These sounds are grouped into the same class that represents a viseme [22]. Different letters can have same visual mouth appearance. In this work Malayalam letters are classified into 16 groups based on Visual Mouth Appearances or visemes ($v_i$) [1]. Lip movements will be almost the same for all letters within a viseme. eg. Letters అ, ఉ, గు (aa, kaa, kshaa) will look same and hence fall into the same viseme. An innovative algorithm has been developed for mapping Malayalam letters to visemes. Detailed test has been conducted for testing the algorithm and the results are verified with typical text data. This has been explained in detail in the thesis.

Emotions can be roughly classified into six categories. joy, sadness, surprise, anger, fear and disgust. This is based on classical work reported by literature [13]. A mixed emotion can be considered as a combination of these basic emotions [16], [26]. So FAP for mixed emotions can be estimated from FAPs for each basic emotion. By using the learning and interpolating property of Neural Networks FAP for mixed emotions can be estimated [19].

FAP estimation involves identification of control points (FCP) on most expressive facial positions and finding the displacement for different visemes and emotions [29]. The FAP has to be normalized. For this facial images of all visemes with neutral emotion and for all emotions without any viseme (for three intensity levels 0.1, 0.5, 0.9) are taken. Taking a neutral face image (no viseme, no expression) as reference, normalized displacements ($d_i$) are found for control points ($P_i$) of each control curve, using vector arithmetic.

For FIS the facial movements for emotions and text has to be combined simultaneously. In this work Three Layer Back Propagation Neural Network together with Digital Image Warping is proposed for FIS. Fig.1.1 shows the scheme.

The software for testing this algorithm is done in C/C++ to run in windows environment. Algorithm was tested on raster and vector images. Results match well with real world talking faces. better on vector images.
all known emotions $\rightarrow e_i \rightarrow$ BPNN (training) FAPs

all known text $\rightarrow v_i \rightarrow$ FAP estimation $\rightarrow$ FCPs of known images

neutral image $\rightarrow$ FCPs of neutral image

emotion $\rightarrow e_i \rightarrow$ BPNN (trained) FAPs $\rightarrow$ Warping $\rightarrow$ sequence of images

text $\rightarrow v_i \rightarrow$

neutral image

Fig.1.1

FIS scheme using Back Propagation Neural Network
1.2 MOTIVATION

Most of the facial animation systems currently available are based on foreign languages. Not much work has been done on Indian Languages. Most of the facial image synthesis is based on speech to facial image synthesis. The speech driven animation system is user dependent. The currently available methods are insufficient for perfect user independent matching of speech signals with visemes. This means every time the voice of the user is changing the system has to be trained. Speech to facial image synthesis incorporating emotions requires identification of emotions from speech. Currently there is no such system which can satisfactorily recognise emotions from user independent speech. For Text to Facial Image synthesis (TFIS) for Indian languages, if English language as a common standard is used, then letters or words have to be mapped to corresponding visemes in English language. Often the mapping procedure has to be innovative as the phonetic characteristics of the English language is much different compared to Indian languages like Malayalam. Taking into consideration of these facts we decided to develop an encoding scheme for classification of Malayalam letters to visemes and make a text to facial animation system for Malayalam Language, combining emotions.

1.3 LITERATURE SURVEY

A text Facial Image Synthesis scheme (TFIS) requires mapping of phonemes to visemes. Several works are already reported. For example English language uses 9 visemes as proposed by Tsuhan Chen and Ram R. Rao [20]. Ashish Verma, Nitenda Rajput, and L.V. Subramaniam of IBM Research Lab in India, Indian Institute of Technology, New Delhi have reported a scheme for mapping phonemes to visemes for Hindi language [22]. In their work they have used Hindi speech recognition system for mapping speech signal to visemes. They have used 12 visemes.

Text to facial Image synthesis with mixed emotions requires detailed study of emotions. In early 1970’s Paul Ekman and his colleagues have performed extensive studies on human facial expressions. They have proposed six basic emotions representing
happiness, sadness, anger, fear, surprise and disgust. They have studied facial expressions in different cultures and preliterate cultures. They have found commonality in the expression and recognition of emotions. They have also observed differences in emotions among various cultures. They have also proposed that facial expressions are governed by “display rules” [13], [14], [15], [24]. Babies seem to exhibit a wide range of facial expressions without being taught which suggests that facial expressions are innate.

Basics of facial animation modelling and synthesis have been studied extensively and several methods have been proposed. One of the methods for synthesizing facial expressions is using 3D wire frame model of facial muscles and texture mapping [25]. [12]. The methods involve measurements of FAPs for different control points on face and interpolation of the position control points on 3D face during synthesizing phase. In their method, they have used 44 basic muscular motions called Action Units (AU) proposed by Ekman and Friesen [16]. They have used different deformation rules of AU for wire frame model of human face. Each facial expression may be defined as a combination of AUs. The coding of facial expressions is done manually by a set of prescribed rules. Their method requires 3D wire frame model of head. The method is time consuming and tedious, and cannot be applied to 2D photo realistic images.

Another method proposed by Turk and Pentland is that decomposing face image into small set of features known as “eigen faces” and Principal Components Analysis (PCA). PCA decomposes a face image into a lower dimensional space known as eigen vectors. The face image can be reconstructed by weighted sum of eigen vectors [53].

Nur Arad demonstrated the use of Radial Basis Functions (RBF) for interpolating anchor points for 2D image warping for synthesizing facial expressions [28]. But their mechanism failed to determine the final position of control points for each particular expression. King and Hou have modified the above technique to estimate the position of control points by considering the six basic emotions. In their method they have estimated FAPs for different degrees of basic emotions, considering them individually. The RBF network was trained using input as intensity of six basic emotions and output as displacements of facial control points [18].

Yan Li, Feng Yu, Ying-Qing Xu, Eric Chang and Heung-Yeung Shum developed Speech Driven Cartoon Animation with Emotions [40]. Their system consists of two
modules. They have developed a method for combining lip-synching method with emotion recognition. In their method they have recorded 1000 utterances of different speakers and Support Vector Machines (SVM) are trained to recognise four emotions neutral, happiness, sadness and fear. They have used audio features such as speaking rate, average length between voice region minimum, maximum and standard deviation of pitch etc are used for training. The hand-drawn sample sketches of face templates of each of the four emotions are also stored. They have also used real time lip synchronisation algorithm for lip synthesis. The audio feature vector MEL spectral coefficients were used for lip synchronization. Finally animation sequence is synthesized by combining cartoon templates with synthesized lips.

S.Morishima, K Aizwa and H. Harashima has developed an intelligent facial image coding driven by speech and phoneme [12]. They have used neural networks for interpolation of 3D facial control points. They have used conventional audio viseme mapping for facial image synthesis. Their method uses linear predictive coefficients (LPC) of speech for input of the neural networks.

1.4 CONTRIBUTION OF THE THESIS

Major contributions of the thesis are listed.

- Developed a scheme for FIS using Neural Network and Digital Image Warping for Malayalam Language.
- Classified Malayalam letters to 16 visemes.
- Developed an innovative algorithm for mapping Malayalam text to visemes.
- Developed a methodology for facial animation parameter estimation to suite the requirement of the thesis.

1.5 ORGANISATION OF THE THESIS

Chapter 2 discusses characteristics of Malayalam language and Classification of Malayalam letters to visemes. A novel algorithm is presented for mapping Malayalam letters to visemes. Test results are also shown.

Chapter 3 discusses the basic emotions and their effects on the face. Encoding of emotion with text information is also discussed. A Text Processor is used for separation of emotion and viseme information.

Chapter 4 reviews FAP estimation. The basis for selection of FCP is also discussed. The structure of FCP is also developed. A suitable methodology for FAP estimation and a novel methodology are developed. Structure of database suitable for neural network training is presented. A software interface is also presented for providing FCP, FAP estimation and preparation of database for neural network training.

Chapter 5 introduces the basis of using neural networks. The various neural network architectures and learning mechanisms are discussed. The BPNN training algorithm is discussed in detail. GUI for neural network training is explained.

Chapter 6 discusses vector and bitmap image formats for facial animation. The vector image modeling using Catmull-Rom spline functions is explained. The porting of standard EPS files to Catmull-Rom spline format is also discussed.

Chapter 7 discusses review of facial image synthesis. Various Digital Image warping methods are introduced and feature based image warping is discussed in detail and it is used in this thesis. Finally the proposed FIS scheme using Back Propagation Neural Network and Digital Image Warping is discussed in detail. GUI for FIS of vector and bitmap images is also presented.

Chapter 8 shows the results of facial image synthesis for bitmap and vector image. Synthesized images are separately shown for individual emotions, visemes, and finally samples are shown for text mixed with emotions.

Chapter 9 concludes the thesis highlighting the merits, drawbacks and further possibilities.

Enclosed CD includes all the software codes in C++, sample input files and output files, and help documentation.