Modeling of 3D Human Face for Tamil Visual Speech Synthesis

V. Anantha Natarajan and S. Jothilakshmi
Department of Computer Science & Engineering
Annamalai University
Annamalai Nagar, India.
friends_mpt2004@yahoo.com, jothi.sekhar@gmail.com

***Abstract* - The visual equivalent of human face exhibiting articulatory expression of a unit of sound in spoken language is termed as Viseme. The Visemes can be used to teach hearing impaired students visually and effectively. In this paper a method to construct three dimensional human faces from a single 2D image is proposed. By this method a generic 3D mesh of human face is deformed using the calculated feature points from 2D images. The method consists of the following steps namely locate corresponding feature vertices in both 2D image and the generic 3D mesh, calculating the corresponding 3D co-ordinates of the 2D feature points and relative camera position and finally deforming the generic 3D mesh using the Radial Basis Function (RBF) based interpolation. Later the constructed three dimensional models can be animated by morphing different visemes together. The inner articulatory organs are modelled using Java3D and blended with these constructed 3D dimensional units.**

Keywords - Viseme, 3D mesh, YCbCr, Radial Basis Function (RBF), KLT feature tracker.

I. INTRODUCTION

Human face modeling is defined as a process by which the two dimensional representations of the human face is translated into three dimensional representation in the form of a mesh which represents the shape of the human face. The computer aided 3D face modelling and animation is not a new subject, but now in recent years there has been increase in the number of researches in this area. This is mainly due to the great advancements and growth of the computation power over the last two decades.

The facial positions and movements, that occurs when voicing a phoneme, is described by a Viseme. Visual Speech can be generated by selecting visemes from a database and blending the units together. The Tamil pronunciation is straightforward in the sense that the pronunciation of each phoneme will not change when they are combined to form a word or sentence. Therefore when they are integrated together to form a visual speech no model of co-articulation is required extensively. Initial step in the visual speech synthesis is the construction of 3D mesh models, each representing a basic sound unit in the Tamil speech. Many different techniques are used for this purpose which may use advanced hardware such as laser scanners, calibrated stereo cameras, or very sophisticated software that can be as expensive as advanced hardware. Face model generation by morphing an initial 3D model using camera image is studied in this research. Automatically extracted feature points on photographs are used to deform initial 3D face model. The 3D face model generation can be viewed as the process of transformation of 2D face image to 3D polygonal mesh representing the face image.

Recent developments in motion sensing technology like the Nintendo and the Microsoft Kinect primarily concentrates on 3D motion capturing for real-time interaction, while it does not care about the geometric accuracy and appearance. This research paper mainly focuses on creating a low cost facial synthesis system that enables arbitrary users to learn a language with a high level of realism.

This paper is mainly focused on generating three dimensional visemes using the feature points taken from single two dimensional images. Since the modelled visemes are used for synthesising Tamil visual speech feature points are lip are considered and extracted using color based segmentation, Haar classifiers and K-means clustering algorithm. Then the three dimensional feature points are calculated by finding the camera calibration parameters and then a three dimensional generic mesh is deformed using the Radial basis function based interpolation method. The section II of this paper describes the various previous works done in the field of three dimensional modeling of human face. In section III the overview of the proposed methodology is presented with results of each task performed during the modeling operation and the section IV concludes the paper.

II. RELATED WORK

An extensive research is going on in the field of computer vision to reconstruct three dimensional objects from a set of images. To construct 3D face models various techniques have been proposed by researchers around the world and one method uses the stereo images (one front view and one side view). Such systems require the users to manually specify the face features on the two images. Generic mesh models comprise of a set of nodes corresponding to characteristic facial points of an average person. The generic model can be incorporated in the algorithm either, before, during or after the 3D reconstruction procedure. In [1] the 3D reconstruction procedure was initialized using the generic mesh model whereas in [2] it was used to constrain the so-called bundle adjustment optimization. In [3], by proposing the incorporation and deformation of the generic model after the optical-flow based estimation of the 3D structure coordinates. The 3D head is parameterized in a statistical way by a shape vector, containing the 3D coordinates of the model's nodes, and a texture vector, containing the corresponding texture value for each node by [4]. Based on interpolating displacements of the boundary nodes a new mesh movement algorithm for unstructured grids is developed by [5] with radial basis functions (RBF's). In [4] using a geometry and an image database a method to construct 3D models of human face from a single photograph is proposed. The system proposed in [4] consumes more computational power. In [6] three dimensional individual faces are constructed by using the extracted information of the facial components such as eyebrows, eyes, nose, ear, mouth and face contour on the face images. Using five orthogonal photographs of the human head and a generic template 3D model a 3D model of a real human head is constructed in [7]. The system proposed in [7] requires manual annotation of the feature points on each face image. By fusing multiple 2D images a 3D face reconstruction algorithm is developed by Changhu Wang, et al., [8]. In his approach initially an efficient multi-view 2D face alignment algorithm is used to locate the facial points and then the intrinsic shape and texture models are extracted by the proposed Syncretized Shape Model (SSM) and Syncretized Texture Model (STM), respectively. In [9] a three dimensional model is developed from a single photograph and various expressions are synthesis using geometric transformations and MPEG-4 facial animation parameters (FAPS). The emerging MPEG-4 standard gives an alternate approach for the researchers to analyse and build models with varying facial expressions [10]. The Language training tools that uses three dimensional facial animations requires that the three dimensional models used for the animation should be accurate and realistic.

III. OVERVIEW OF THE MODELING METHOD

In this research the images are taken from the video frames of the Tamil TV anchor. The images are taken without any calibration constraint and in this study four feature points are used for model generation. To find the 3D coordinates of a feature point it is important that the feature point must be located in at least 2 different frames. In each 2D video frame all the four feature points must be specified. Then using an iterative algorithm the 3D coordinates of the feature points are calculated. The next step is deforming a generic face mesh considering the coordinates of the 3D feature points found in the iteration phase. This less quantity of feature points will shape the rest of the face by employing a Radial Basis Functions. The Figure. 1 below shows the step by step tasks involved in modeling a three dimensional human face.

After deformation, if results are not satisfactory, then some more new refining feature points can be added to the initial images. This sub-step of deformation is called refining in this work. The aim of the refining step is to specify more feature points which are specific to the current subject's face. There are two reasons why we do not include these refining feature points in camera calibration. First, these points may not be easily marked on images and can affect calibration of camera. Second, these 3D feature points may be specific to the current subject only and may not be marked on the initial generic 3D face model.

The last step is forming of a texture image to wrap the final 3D face mesh. The ultimate aim of our research is building a speech training tool, the main requirement of the speech training is that all the articulatory parts must be visible and if the 3D models are textured with human skin maps then the position of inner articulatory parts will not be visible. So a simple transparent skin is applied to the 3D model.

A. Feature Point Detection

In the first frame of the video the feature points around the lip region is found automatically using a contrast stretching technique then using the KLT feature tracker algorithm the feature points in the other video frames are tracked. For finding the feature points in the initial frame of the video the face region is localized using the Skin color based segmentation in the *YCbCr* color space, then the lip region if found using HAAR classifier. After finding the region of interest the contrast is stretched resulting in a high contrast sub-image. Then the lower and upper limit of the pixels in the image is found. Then shift and scale the pixels values so that limits are scaled between 0 and 100%. In a contrast stretched image the lip region alone will be highlighted then the edge is detected which gives the lip contour. And finally using a line scanning algorithm the left and right corner points are detected. Then the centroid of the corner points is calculated. Then the pixels with minimum and maximum y-coordinates in the x-coordinate are found which is the top and bottom feature points of the lip contour. Totally four feature points around Lip contour are detected, if the deformation of the generic mesh is not smooth more feature points can be added. The Figure 2. shows the result of the face and lip localization using skin color based segmentation and Haar classifiers respectively [11]. The methods used to detect face region and localize the lip in this research is more reliable and Robust.

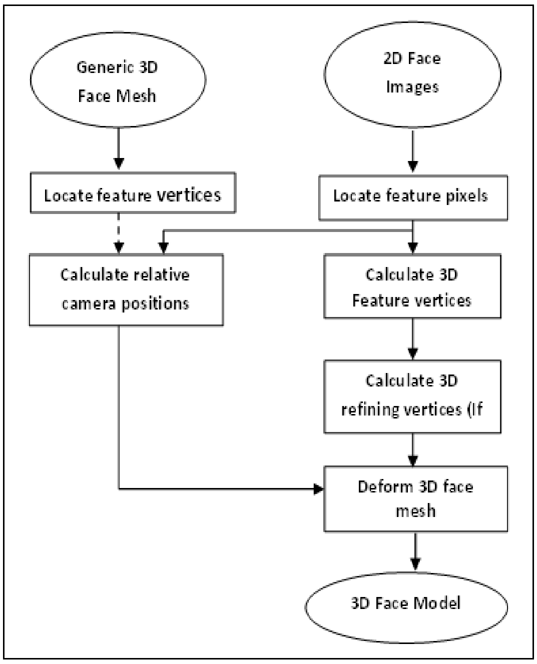

Figure 1. Block Diagram of the proposed three dimensional modelling

The step by step result in finding the lip contour in the extracted lip region is shown in Figure. 3. The lip region is contrast stretched and converted in to binary. Using canny edge detection algorithm the lip contour is detected. The detected lip contour is verified manually on all the initial frame of the collected videos which is found to be more accurate and appropriately suitable for detecting the feature points around it.

Finally using the scanning algorithm the left and right corner points are detected and then the top and bottom feature points are also detected. A scanning algorithm is chosen since the area to be searched for the corner points in smaller and it is computationally less expensive.

B. 3D Feature Point Estimation

3D Feature point estimation is done in three steps which are camera calibration, 3D Coordinate calculation and data re-normalization. Camera calibration is the process of calculating the perspective projection matrix of the camera which converts the 3D real world coordinates to a 2D image and it is expressed as follows:

$$s\begin{bmatrix} x \\ 1 \end{bmatrix} = P\begin{bmatrix} X \\ 1 \end{bmatrix} \qquad (1)$$

where x = (x,y) is 2D image coordinates and X = (x,y,z), is 3D coordinates. **P**, which expressed in homogeneous coordinates, is a **3x4** camera projection matrix and is the arbitrary scale factor. Eq. 1 can be rewritten as:

$$S\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & 1 \end{bmatrix} \begin{bmatrix} X_i \\ Y_i \\ Zi \\ 1 \end{bmatrix} \qquad (2)$$

For each 2D feature point we have two equations:

$$u_i = \frac{X_i P_{11} + Y_i P_{12} + Z_i\, P_{13} + P_{14}}{X_i P_{31} + Y_i P_{32} + Z_i\, P_{33} + 1}$$

$$V_i = \frac{X_i P_{21} + Y_i P_{22} + Z_i\, P_{23} + P_{24}}{X_i P_{31} + Y_i P_{32} + Z_i\, P_{33} + 1} \qquad (3)$$

To find **P** we rearrange these equations as follows:

$$X_i P_{11} + Y_i P_{12} + Z_i P_{13} + P_{14} - u_i X_i P_{31} - u_i\, Y_i P_{32} - u_i Z_i\, P_{33} = u_i$$
$$X_i P_{21} + Y_i P_{22} + Z_i P_{23} + P_{24} - \ V_i\ X_i\ P_{31} - V_i\ Y_i\ P_{32} - V_i\ Z_i\ P_{33} = v_i \qquad (4)$$

The feature points are extracted on all frames and the generic 3D mesh model. To find the 3D coordinates of feature points, the two dimensional coordinates of the extracted feature points from the video frames and calibration matrices from the video frames is used as shown below:

$$X\left(p_{11}^i - u_i\, p_{31}^i\right) + Y\left(p_{12}^i - u_i\, p_{32}^i\right) + Z\left(p_{13}^i - u_i\, p_{33}^i\right) = u_i - p_{14}^i$$
$$X\ \left(p_{21}^i - v_i\, p_{31}^i\right) + Y\ \left(p_{22}^i - v_i\, p_{32}^i\right) + Z\ \left(P_{23}^i - v_i\, p_{33}^i\right) = \ v_i\, P_{24}^i \qquad (5)$$

where *u* and *v* are two dimensional coordinates of the extracted feature points on the i^th frame and *p*’s are the entries of the perspective projection matrix (P) of the same frame. To have a valid solution for this equation each feature point extracted from the frame must be specified in at least two successive frames. The Eq.5 is rewritten in to linear equation form and solved using pseudo-inverse method.

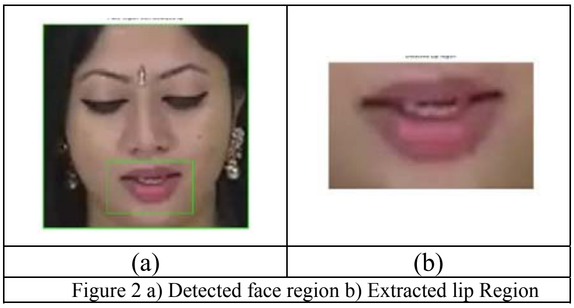
(a) (b)

Figure 2 a) Detected face region b) Extracted lip Region

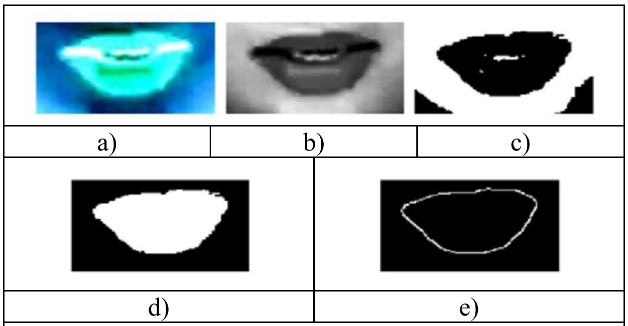
a) b) c) d) e)

Figure 3. a) Contrast stretched Lip region b) Gray image c) Lip region in Binary d) masked Lip region e) Edge detected

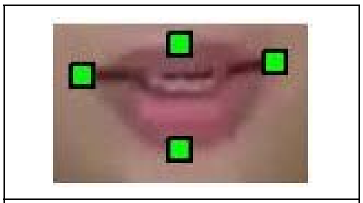
Figure 4. Detected Feature Points

C. 3D Face Deformation

Feature points are used to deform all the vertices in the generic face mesh. A smooth interpolation function will be constructed for the 3D displacements of the mesh vertices. Given *n* feature points, displacement for a feature point *i* is defined as follows:

$$d_i = X_i - X_i^0$$

where X_i , X_i^0 are final and initial 3D coordinates of i^{th} feature point.

Radial Basis Function (RBF) interpolation is used to interpolate globally between samples scattered in the n-dimensional space. The huge number of RBF's available and their various parameters, make the RBF interpolation very flexible. We used gaussian function in our implementation and it is denoted by,

$$\text{Gaussian function} \quad G\,(d,\varepsilon) = c^{-(d\varepsilon)2} \qquad (6)$$

D. Modeling Inner Articulators

Visual speech training can be successful only if the constructed 3D models reflect the actual articulatory organs in its respective positions. The articulatory positions of various sound units in Tamil language are listed in table below. Modeling these internal articulators are not so easy by acquiring knowledge from the 2D images so the position of these articulators were collected from the various resources and individually modelled using the Java3D API. Using the API gums and tooth's are modelled using the polygon cubes and it is textured with teeth image. Next the tongue is modelled using the smooth polygon cube. For each position of tongue a separate model is constructed. The Table 1 presents the position of articulators for each phoneme pronunciation.

IV. RESULTS & DISCUSSION

In Figure 4, the generated three dimensional models for different Tamil phonemes are shown. The side view is chosen since it better displays the position of the inner articulators. A GUI based user interface has been developed to create three dimensional visual representations of different Tamil speech units and when the user enters the text in Tamil it will be syllabified and the corresponding visual representations stored in the form of OBJ models will be taken Morphed to create the visual speech. To control the smoothness in the animation the timing information extracted for the syllables during speech syllabification is utilized. Each key frame needed for the animation is extracted from the stored OBJ models corresponding to the given text input.

Once the synthetic faces are modeled, to generate visual speech synthesis the models have to be animated. As previously mentioned, we have used key-frame approach of animation.

TABLE I. ARTICULATORY POSITIONS FOR TAMIL PHONEMES

#	A	LIP	TONGUE
1	m	The lips are kept in neutral positions	Tongue is kept as it is in the normal position
2	,	The lips are spread	The front of the tongue is raised as high as possible towards the hard palate. The tip of the tongue touches the lower part of the teeth
3	C	close lip rounding	The sides of the tongue are in contact with the teeth and the tip of the tongue is somewhat retracted from the lower part of the teeth.
4	v	The lips are slightly spread	The front of the tongue is slightly raised towards the hard palate but not as high.
5	I	The lips are kept in neutral positions, the lips are spread	Tongue is kept as it is in the normal position the front of the tongue is raised as high as possible towards the hard palate. The tip of the tongue touches the lower part of the teeth.
6	x	The lips are rounded with considerable protrusion	The back of the tongue is raised towards the soft palate.
Key to column headings: # = Serial No., A = Alphabet			

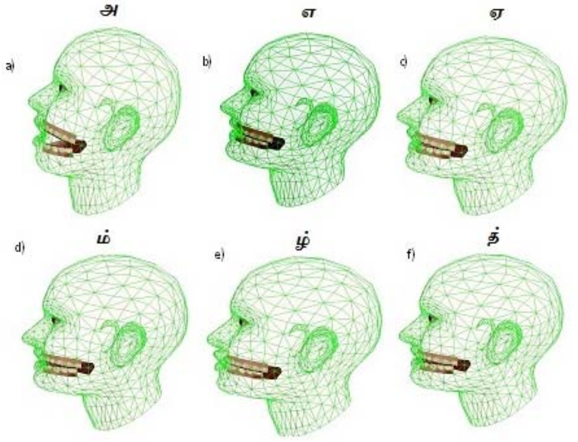

Figure 4. Three dimensional models for Tamil speech sounds

V. CONCLUSION

The face model generation by deforming a generic 3D model using un-calibrated camera image is performed with some simple problems. On the video frames feature points marked automatically using the proposed method are used to deform initial 3D face model. Since the models generated are used for synthesising visual speech for training the hearing impaired students the focus is lip region, no attention is paid for synthesising expressive speech. In our future work the research will be focussed towards modeling the 3D face models with facial expressions. The creation of facial synthesis is normally done by sculpturing keyframe faces for every two or three frames. When more number of frames is to be created Facial motion capture system is widely used to acquire human facial motion data, which is quite expensive. In this research human facial motion data is extracted automatically from the video frames and these data are used to model three dimensional human face. Extracting these data points accurately is a tough task to implement with limited resources. In future if a motion capture system is utilized to record the human facial motion data, the complexity in developing a facial synthesis system could be simplified.

REFERENCES

[1] S. Basu A. Pentland J. Strom, T. Jebara. Real time tracking and modeling of faces: An ekf-based analysis by synthesis approach. Technical Report, page 506, MIT Media Laboratory, 1999.

[2] Z. Zhang Y. Shan, Z. Liu. Model-based bundle adjustment with application to face modeling. In Proc. of IEEE Internation Conference on Computer Vision (ICCV), volume 2, pages 644–651, Vancouver, BC, Canada, 2001.

[3] S. Krishnamurthy T. Vo A.R. Chowdhury, R. Chellappa. 3d face reconstruction from video using a generic model. In Proc. of IEEE International Conference on Multimedia and Expo (ICME), volume 1, pages 449– 452, 2002.

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In Proc. of the International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), pages 187–194, New York, NY, USA, 1999.

[5] De Boer A, van der Schoot MS, Bijl H. Mesh deformation based on radial basis function interpolation. Comput Struct ;85(11–14): 784–95.

[6] SunHee Weon, SungIl Joo, and HyungIl Choi. Individualized 3D Face Model Reconstruction using Two Orthogonal Face Images. In Proc. Of the World Congress on Engineering and Computer Science volume 1, San Francisco, USA, 2012

[7] G. Fanelli, M. Fratarcangeli. A non-invasive approach for driving virtual talking heads from real facial movements. 3DTV Conference, pages 1-4, 2007.

[8] Changhu Wang; Shuicheng Yan; Hongjiang Zhang; Weiying Ma. Realistic 3D Face Modeling by Fusing Multiple 2D Images. In Proc. Multimedia Modelling Conference, pages 139 – 146, 2005.

[9] N. Patel and M. Zaveri. 3D Facial Model Construction and Expressions Synthesis using a Single Frontal Face Image. International Journal on Graphics, volume 1, 2010.

[10] Y. Zhang, Z. Zhu and B. Yi. Dynamic facial expression analysis and synthesis with MPEG-4 facial animation parameters. IEEE transactions and systems for video technology, volume 18, no. 10, pages 1383-1396, 2008.

[11] C. Erdem, S. Ulukaya, A. Karaali and A. Erdem. Combining Haar Feature and skin color based classifiers for face detection. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011.