© IADIS, Proceedings of the IADIS International Conference on WWW/Internet , October 2006, Murcia, Spain, pp. 154-161

LOW BANDWIDTH VIDEO CONVERSATION USING ANATOMICAL RENCONSTRUCTION OF FACIAL EXPRESSIONS OVER THE INTERNET

Arvind K. Bansal and Y. Hijazi
Department of Computer Science
Kent State University, Kent, OH 44242, USA
E-mail: arvind@cs.kent.edu, Ph: +1.330-672-9035

ABSTRACT

As the Internet becomes fast and ubiquitous, the demand for realistic multimodal communication will increase substantially for human-like human computer interaction and video conversation over the Internet. This demand will easily undo the bandwidth gain of the Internet2. This will necessitate realistic image visualization and rendering at the client end using a low bandwidth textual description of dynamic changes in multimodal interaction. In this model we describe an integration of a client end detailed anatomical model of facial expressions superimposed on a single high resolution image of a speaker. The facial expression rendering changes dynamically based upon textual information about emotional expression transmitted over the Internet. This paper describes FEML (Facial Expression Markup Language) — an XML based language for 3D dynamic anatomical modeling of human face expression over the Internet and its integration with the corresponding sound transmitted over the Internet. FEML has two parts: the user level that encodes the emotions and the corresponding audio, and the system level macros for anatomical muscle modeling of facial expression. The performance evaluation shows that for longer conversation the efficiency achieved is significant to frame based audio-visual transmission formats such as MPEG movies.

KEYWORDS: Bandwidth, Internet, Language, Video conversation, facial expression, XML

1. INTRODUCTION

Considerable interest has developed in computer-based three-dimensional facial character animation. The initial effort to animate and represent the face using computers goes back to the last three decades. However, in the recent years, due to the availability of multimedia communication over the Internet and inexpensive desktop computer processing power, realistic facial modeling for communication has become more desirable. Recent interest in facial modeling and animation has been spurred by the increasing appearance of virtual characters in film and video, and the potential for a new 3D immersive communication metaphor for human-computer interaction such as haptic faces [12]. Video phoning and video conference is becoming quickly a desired means of communication both for cost cutting, job efficiency, and real time collaboration. However, the real bottleneck in the real time video collaboration and video helpdesks is caused by amount of data needed to transmit 30 frames per second. If a reasonable size colored image of 400 X 240 with 20 K bytes/frame for a compressed MPEG image is transmitted that will amount to 200 MB of data for a five minute video conversation including audio. Assuming 1000 transmissions involving video clips, video conversations, and multimedia movie transmission at a time over a local residential area network, there will be a load of 1GB/sec sufficient enough to overload the capacity of the Internet2. This makes it necessary that we complement the increase in bandwidth with the reduction in bandwidth requirement. While image compression like MPEG-4 will help the cause significantly, the problem of video conversation can be helped realistically by dynamic modeling of human emotions at the client end, and transmitting the sound over the Internet.

Client end modeling of 3D images to reduce the bandwidth requirement is a well researched concept [9, 10, 16, 23, 26, 27], and has been researched by the first author for transmitting movies over the Internet using STMD (Single Transmission Multiple Display) model [23]. In STMD model, a 3D image is made of multiple components, and the components are cached and reused to build images at the client end. In this research, we extend the notion of STMD model to automatically generate facial expressions (at the receiver end) for realistic video conversation over the Internet. The basic idea is to represent the face as a multilayer, transmit the emotional expression change, textual or audio speech and the emotional intensity over the Internet, simulate the anatomical model of face at the client end, and integrate the audio with the anatomically generated facial expression. The facial expression is modeled using a generic 3D triangular mesh for the 3D bone structure and skin modeling, anatomical muscles are modeled using a combination of trusses and springs, facial expressions are generated using anatomical muscle displacement techniques [1, 3, 4, 5, 17, 18, 21, 24, 28, 29, 30, 31, 32], and morphed image is superimposed on the top of the 3D mesh structure for realistic face animation. We also model the transition of the dynamic change of facial expression and the corresponding voice using finite element method [8]. The overall architecture is described in Figure 1.

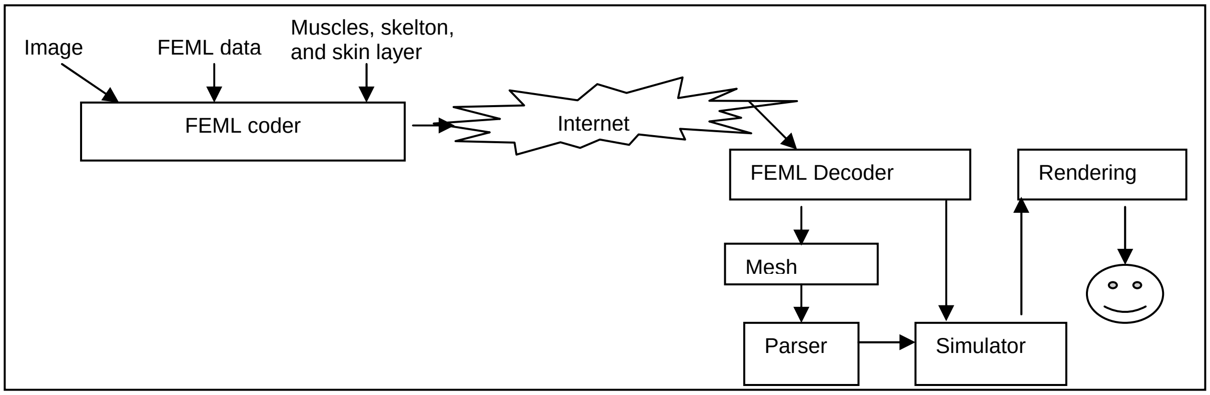

Figure 1. The overall model

The voice is categorized into multiple emotional phonemes — a set of phonemes variant to take care of individual variations of different phonemes according to emotions. At the sender end, the textual description of emotional sound is transformed automatically to an XML format incorporating emotions, corresponding text is associated with the given emotion, and transmitted to the client end along with the periodic face image taken from camera or archived in the database. At the listener end, the XML file is decoded. First the associated image (if any) is superimposed on the 3D mesh structure, then XML file is transformed to muscle movement macros needed for facial expressions, and text to emotional speech conversion is performed by indexing the emotional phonemes to derive the corresponding muscle movements. Facial expression synthesis is achieved by modeling different combinations of a set of contractions of facial muscles as described in FACS [4, 5].

The major contributions described in this paper are: (1) a new XML based language has been developed for a remote simulation of facial expressions, (2) due to the sparing use of the transmitted image the data transfer overhead has been reduced by an order of magnitude while maintaining the realistic quality, (3) the talking face image can be superimposed with other images dynamically, and (4) the emotion can be altered dynamically.

The system has been implemented [8] using C++, XML, and Microsoft Speech SDK library [14]. This system supports media objects such as background, text, images pf real people, and audio.

The paper is organized as follows. Section 2 describes the background and the related definitions. Section 3 describes the model architecture, Section 4 describes a BNF grammar for FEML. Section 5 describes the performance evaluation. Section 6 describes the related works. Section 7 describes the limitations and future work, and the last section concludes the research.

2. BACKGROUND

In this section, we briefly describe the background concepts needed to simulate the three dimensional facial expressions using anatomical models of muscles, different approaches for automated modeling of facial expressions, and basic action units for modeling the facial expressions using anatomical muscle based modeling.

A *normal stress* is either compression or tension. *Shear* is a type of stress that acts parallel to a surface, and deform objects by rotating the axial lines. *Strain* is defined as the fractional deformation of an elastic object under stress. A *truss* is a slender member that can support an axial load, and is capable of tension and compression. *Tensors* are arrays of functions that transform according to certain rules under a change of coordinates.

Finite element analysis is based upon dividing a problem domain into many interconnected sub-domains called finite elements. The differential equations are solved using computer based numerical techniques for each finite element, and then linked to adjacent finite elements using boundary conditions. Increasing the number of finite elements improves the accuracy at the cost of computational efficiency. Finite element analysis uses a complex system of points called *nodes* that make a grid called a *mesh*. A mesh is programmed to model the reaction of a structure to loading conditions. Nodes are assigned at a density throughout the material depending on the anticipated stress levels of a specific region. Regions that receive large amounts of stress usually have a higher node density. Each node has associated vectors that carry the material properties to adjacent nodes.

The soft tissue of the skin is visco-elastic in its responses to stress [24, 30]. However, the relationship between load and deformation is nonlinear: Under low stress, skin tissue offers low resistance to stretch, but under greater stress it resists more.

A *phoneme* is the basic unit of the sound and the corresponding shape of the mouth and lips is called a *viseme*. There is one to one correspondence between a phoneme and a viseme. There are three major approaches to model animated facial expressions in response to phonemes: viseme to viseme transitions [2], automated modeling of muscles [1, 3, 4, 5, 21, 22, 24, 28, 29, 30, 31, 32], and hybrid approaches integrating image morphing and geometric 3D modeling [20]. While viseme to viseme transition and hybrid models are computationally efficient, they do not produce realism due to limitations imposed by finite number of available combinations. In contrast anatomical muscle modeling produces more realism.

In anatomical muscle based modeling of facial expressions, all muscles have an origin and are attached to a specific location on the skull. There are three types of muscles: linear, sheet, and sphincter. Linear muscles are one dimensional, sheet muscles are a group of fibers that stretches non-uniformly along an axis in two dimensions, and sphincter muscles are ring shaped (such as mouth) that enlarge or reduce by relaxation or compression. FACS [4, 5] is a set of 66 possible basic actions performable on the human face. An *action unit* (AU) is a single action caused by the movement of a subset of closely related facial muscles. A small subset of closely related action units simulates a facial expression. For example, combining the AU_1 (inner brow raiser), AU_4 (brow raiser), AU_{15} (lip corner depressor), and AU_{23} (lip tightener) creates a sad expression. Tables I and II illustrate a sample set of required action units for generating facial expressions.

Table 1. Sample single facial action units

AU	FACS name	AU	FACS name	AU	FACS name	AU	FACS name
1	Inner brow raiser	6	Cheek raiser	12	Lip corner puller	17	Lower lip depressor
2	Outer brow raiser	7	Lip tightener	14	Dimpler	20	Lip stretcher
4	Brow lower	9	Nose wrinkler	15	Lip corner depressor	23	Lip tightener
5	Upper lid raiser	10	Upper lip raiser	16		26	Jaw drop

Table 2. Different expressions and action units

Basic expressions	Involved action units	Basic expressions	Involved action units
Surprise	1, 2, 5, 15, 16, 20, 26	Sadness	1, 4, 15, 23
Fear	1, 2, 4, 5, 15, 20, 26	Happiness	1, 6, 12, 14
Disgust	2, 4, 9, 15, 17	Anger	2, 4, 7, 9, 10, 20, 26

3. MODEL ARCHITECTURE

The geometric manipulation of a face needs the construction of a fully functional 3D model of a subject's face using triangular facial mesh. The 3D facial model incorporates a skull structure that allows us to set the facial muscles in accurate positions, and connect the muscles, fat tissue, and skin on an actual face. The coordinate of the nose is used for the initial alignment of 3D skull mesh and the user transmitted 2D image. The other fixing points needed to align the specific part on 3D skull to 2D image during superimposition are: (i) head top (ii) the jaw joints, (iii) chin as a movable fixed point due to jaw movement, (iv) four fixed point defining a nose, and (v) four fixed points defining each eye, and (vi) the upper jaw point. For realistic modeling of muscles, all major anatomical muscles were simulated as a combination of single dimension trusses, two dimension trusses, and a network of springs (to model fat) with proper fixation points on the skull modeled as a 3D mesh. Based on the distribution of muscles, the face has been divided into five mutually exclusive regions: cheek (right and left), forehead (right and left), on either side of the nose, mouth (upper lip, lower lip mouth corner), eyes (right, left), eyebrow and eye lid. The reason for dividing the face is to (1) control and limit the displacement that is produced by muscle contractions, and (2) to reduce the computational overhead of finite element method.

Skin is modeled as a 3D mesh on top of the muscles as shown in Figure 3. The synthetic skin tissue is modeled as a set of polygonal elements that are divided into the skin surface and the facial tissue surface. Facial tissue surface and skin surface are separated by fat layer. Springs connect the skin and the facial tissue layers to simulate skin layer elasticity. The tissue model is composed of triangular mesh elements that match the triangles in the adapted facial mesh. Muscles are fixed in a bony subsurface at their point of emergence (referred to as 'origin') and are attached to facial nodes (denoted as 'end points') as they run along tissue elements.

Upon the application of a force the corresponding muscles contract. As the muscles contract the corresponding mesh points also change their positions. This causes the facial skin points (mesh points) that are in the influence area of the muscle to be displaced to their new positions due to sheer. The movements of skin points also affect neighboring skin points in the mesh based upon a pre-defined weight matrix based upon the connectivity and direction of the displacement of the mesh points.

Talking face is modeled as a multi-frame movie. Multiple frames are automatically generated at the client end after superimposing the 2D image transmitted by the speaker on a 3D skeleton at the listener end. At the user level the speaker only describes the two dimensional image of the actor. Combining the contractions of facial muscles using FACS creates various facial expressions. After the initial frame is generated, the following frames related to the same emotions are generated by transforming the change in emotional intensity that uses tensors and finite element method on the corresponding facial regions to model the displacement in response to the required forces. Higher intensity increases the forces on muscles proportionately.

Text to animated speech conversion is based on synchronizing the automatic generation of visemes and facial expression. After receiving the sequence of phonemes, viseme table is looked up for the corresponding mouth deformation. Lip movement for animated speech [1] is produced by interpolating between the visemes to produce a series of frames in between a start picture and an end picture generated by anatomical muscle modeling [29]. The two processes of phoneme to sound conversion and the visualization of the remaining facial expression are launched simultaneously to synchronize the face movement with the sound. By knowing beforehand the total length of the sound, the number of frames of the animation (given by the temporized visemes), and the current time, the synchronization is done by skipping frames in case of delay in the rendering process (see Figure 4).

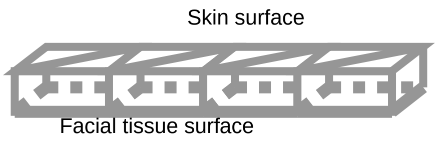

Figure 3. Modeling skin and muscle interface

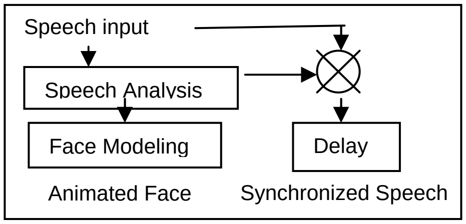

Figure 4. Synchronized animated talking

4. FEML – FACIAL EXPRESSION MARKUP LANGUAGE

In this section we give an Extended BNF form of the grammar for FEML. For the convenience of readership, we have omitted low level XML syntax or DTD.

The <*Talking-Face*> tag is the root of the grammar, which contains all the needed attributes and one or more emotional talk sequence. The <*Face-attr*> includes <*Frame-rate*>, <*Size*>, <*Location*>, *and* <*Style*> of the window space in which the clip is rendered. The <*style*> attribute takes on either of two values: "window" (default mode) or "full screen". The <*text*> tag is a sequence of <*characters*> for interaction with the listener or to provide as a caption below the image of the actor to identify him.

Emotional talk sequence <talk-seq> consists of one or more talks. Each talk may correspond to a single emotion during which a new two dimensional image is transmitted and superimposed on initially transmitted three dimension generic facial model.

A <*Talk*> includes emotional-state <*State*>, <*Actor-image*>, <*Speech*> tags. An <*Actor-image*> consists of image attributes and the name of the image file. An emotional state could be a <*Simple-emotion*> or a <*Blended-emotion*>. A <*Blended-emotion*> is comprised of multiple <*Simple-emotions*> mixed in different weights <Weight>. The tag <*Emotion-type*> could by user-defined blended emotion name or one of the base emotions such as *happy*, *sad*, *surprise*, *passive*, *angry*. The tag<*Intensity*> describes the strength of emotion as an integer value. The tag <*Speech*> could be either a transmitted text file <*Text-speech*> or an audio file <*Audio-speech*>. The tag <Text-speech> represents a synthetic speech reconstructed at the client end using library of phonemes of the sender. The tag <Audio-speech> transmits the actual audio file <Audio-file> to the client end.

The talking face can have one or more <*Actor*> tags embedded within it. The tag <*Actor*> describes the name of an actor and the corresponding image file <Image>. The tag <*Location*> is split up into having two subparts: <*url*> and < *Image-id*>. The <*url*> gives the path on the server, and the < *Image-id*> acts as an index to transmit the corresponding 3D generic mesh and the face model for the corresponding actor.

```
<Talking-Face> :: =‘<Face’ <Face-attr>’>’<Actor>
          {<Talk-seq >}+  ‘</Face>’
<Face-attr> ::=   [<Frame-rate>] [<Size>] <Location>
   [<Style>] <Caption>
<Frame-rate> ::= ‘fps =’ <integer>
<Style> ::= ‘style=’ <style-options>
<Style-options> ::=  <Size> | ‘fullscreen’
<Location> ::=  ‘location =’ <url>’/’<Image-id>
<Image-id> := <Name>
<Caption> ::=  ‘caption =’ <text>
<Size> ::= ‘size =’ ‘(‘<Integer>’,’<Integer>’)’
<Actor> ::= ‘<actor>’ <Name> <Image>   ‘</actor>’
<Image> ::=  ‘<image’  <Size>’>’<Image-file>
                 ‘</image>’
<Image-file> ::= <Name>.<Image-format>
<Talk-seq>::= ‘<talk>’{<Talk>}*‘</talk>’
<Talk>::= <State>[<Image-file>]<Speech>
<State> ::=  <Blended-emotion> | <Simple-emotion>
<Blended-emotion> ::= ‘<Blended-emotion’
                <Emotion-attr> ‘>’{<Simple-emotion>
                <Weight>}+  </Blended-emotion>

<Simple-emotion>::=  ‘<emotion’<Emotion-attr>‘>’
<Emotion-attr> ::= <Emotion-name> [<Intensity>]
<Weight> ::= ‘<weight>’ <Integer> </weight>
<Emotion-name> ::=  ‘Emotion-name =’ <Emotion-type>
<Emotion-type> ::= <Name> | <Base-emotion>
<Base-emotion>::=‘happy’|‘sad’|‘relaxed’|‘surprised’|
                          ‘angry’| ‘upset’|‘disgusted’
<Intensity> ::= ‘intensity =’ <integer>
<Speech>  :: = <Text-speech> | <Audio-speech>
<Text-speech>  :: = ‘<text-speech’ <Speech-attr>’>’
                                <Text> ‘</text-speech>’
<Speech-attr>  ::= [<Sample-rate>][<Amplitude>]
<Audio-speech> ::= ‘<Audio-speech>’ <Audio-file>
                           ‘</Audio-speech>’
<Text> ::= {<Character>}+
<Audio-file> ::=  <Name>‘.’<audio_format>
<Audio-format> ::= ‘avi’|’wav’
<Image-format> ::= ‘jpg’|’gif’
<Sample-rate> ::= ‘sps =’ <integer>
<Amplitude> ::= ‘amplitude =’ <integer>
<Name> ::= {<Character>}+
```

Figure 5. User level Extended BNF grammar for FEML

The user level XML specification is translated to system level macros at the client end that also include muscle displacements to simulate facial expressions, 3D mesh for skull representation, and 3D mesh for skin representation. Interested readers may find the system level details in [8].

5. PERFORMANCE EVALUATION

The XML clips files used in these experiments set the number of frames per second to 30. Since different regions have different 3D mesh complexity, the facial expressions involving more number of triangular meshes required more time. For our study, we took a face with forehead having 312 vertices, eyes having 798 vertices, nose having 267 vertices, cheek having 159 vertices, and mouth region having 944 vertices.

The rendering time includes time taken to parse XML file, finite element simulation, time to retrieve visemes for the corresponding phonemes, and the time taken to render the composite face. Figures 6, 7 and 8 show the performance measurement of the implemented system on two desktop PCs with 1 GB of memory and a Pentium 4 processor with 2.4 GHZ clock speed and 533 Mb/sec memory bus connected to the Internet through a 10 MB/sec network local area network.

Figure 6 plots the time taken to model and render different facial regions. Since different emotions require different combinations of regions, the rendering time for different emotions vary. For example, rendering the first instance of fear takes 2546 milliseconds, rendering the first instance of surprise takes 2000 milliseconds, rendering the first instance of happiness takes 1238 milliseconds, rendering the first instance of sadness takes 1326 milliseconds, rendering the first instance of disgust takes 1549 milliseconds, and rendering the first instance of anger takes 1625 milliseconds. Figure 7 compares the amount of data transfer overhead needed to play the same clip in our system with pre-archived MPEG movies transmitted over the Internet using a regular load. The X-axis shows the clip duration, and the Y-axis shows the data transfer overhead. The result show that for small clips (less than five seconds), the initial overhead of rendering in muscle based systems is large enough so the advantage of client end rendering of facial expressions is not significant. However, as the clip time increases beyond 60 seconds, the muscle based modeling system requires much smaller amount of memory, and ratio of MPEG based clips and muscle based modeling increases significantly. Figure 8 shows the overhead of the initial time taken to update the state of all the regions of the faces based on their default animation state. The time taken increases with the resolution of the clip in the movie. The initial overhead is around 25 milliseconds for an 800 × 600 image. After this initial time, the time taken to update the state of a face based on the facial expression is around 15ms. This is due to the fact that the faces and libraries were cached and accessed during the initial setup.

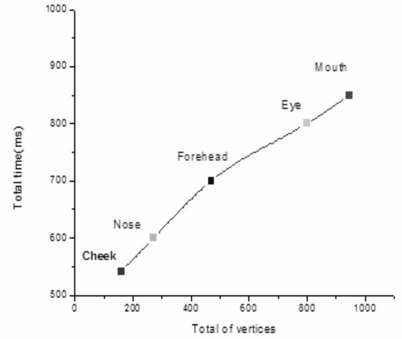

Figure 6. Rendering time for facial regions

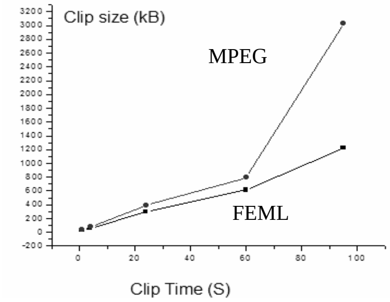

Figure 7. Data transferred vs. clip duration

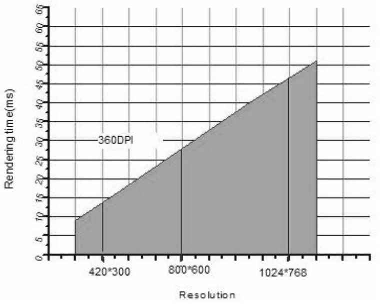

Figure 8. Initial rendering time vs. image resolution

6. RELATED WORKS

Since the pioneering work of Parke [18] in 1972 and later by Ekman and Friesen [4] to abstract the muscle movement using FACS and vector based muscle model by Waters for 3D facial animation [28], many researchers have attempted to generate realistic facial modeling and animation. The first major work to integrate speech with facial expression was done by Waters and others [29, 30]. Facial modeling and animation research falls into three major categories: geometric based modeling involving parameters [15], anatomical muscle based modeling [3, 4, 5, 13, 19, 24, 28, 29, 30, 31, 32] and image morphing [2, 20].

We have been influenced mainly by image based morphing and anatomical muscle based modeling. Our model integrates the features from both the categories for an optimum result. We use image manipulation model for visemes and integration of muscle based modeling and finite element method for facial expressions.

We have incorporated truss based modeling for non homogeneous muscle deformation of sheet muscles. Our approach uses matrices and non-linear equations to model non linear elasticity of muscles, their constrained motions based upon their attachments on the skull, and non linear elasticity in skins. To the best of our knowledge truss based approach and handling of non-homogeneous deformation, automatically generated facial expression for Internet based video conversation, the development of 'Facial Expression Markup Language' (FEML) and its integration with automated facial expression generation has been attempted for the first time.

Another related work is the development of high level languages Mimic to simulate the facial expressions programmatically [6]. This work is quite different in the motivation, as we are interested in Internet based transmission of video conversation. However, mimic can be integrated to FEML for Internet based transmission for remote human-computer interaction over the Internet.

7. CURRENT LIMITATIONS AND FUTURE WORK

The current implementation does not exploit the knowledge of the variable thickness of skin layers in different area of the face and the formation of wrinkles [24, 31], head and neck movement such as shaking the head during emotional laughing, and change of voice with modes of a speech such as r*equest, inquisition, command, sarcasm, fatigue, excitement, boredom, amusement* etc. These modes of speech are largely independent of the emotions, and need to be superimposed with emotion. Another limitation is that the voice of a person is transmitted once and stored in the library at the listener end, and the voice is reconstructed using archived phonemes. A better way would be to dynamically archive the realistic phonemes along with the emotional intensity in real time, and use that phoneme for voice construction at the receiver end. The fixing points were manually analyzed to align 2D image with the 3D skull mesh. In future implementation image analysis technology already present in literature will be used for an automated analysis.

8. CONCLUSION

In this paper, we have described an Internet based system for the low bandwidth video conversation over the Internet. The system uses receiver end reconstruction of facial expressions and the use of visemes corresponding to emotional phonemes. The results show that the system can significantly reduce the transmission overhead compared to popular audio-visual formats such as MPEG for longer version of conversation while maintaining the picture quality and image. However, the realism is still restricted for finer details such as voice inflections caused by different modes, incorporation of facial wrinkles, and finer changes in facial expressions during emotional changes. In future, we intend to improve these features.

REFERENCES

[1] Basu, S., Oliver, N., Pentland, A., 1998. 3D Lip Shapes from Video: A Combined Physical-Statistical Model *Speech Communication,* Vol. 26, pp. 131-148.

[2] Beier, T., Neely, S., 1992. Feature-based Image Metamorphosis. *Computer Graphics,* Vol. 26, Issue 2, pp. 35-42.

[3] Bui, T. D., Heylen, D., Poel, M. and Nijholt, A., 2003. Improvements on a Simple Muscle-based 3D face for Realistic Face Expressions. *Proc. 16th Int. Conf. on Computer Animation and Social Agents.* New Brunswick, USA, pp. 33-40.

[4] G. Donato, M. S. Bartlett and J. C. Hager et. al., 1999. Classifying Facial Actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 21, No. 10, pp. 974-989.

[5] Ekman P. and Friesen, W., 1978. *Facial Action Coding System Manual*. Psychologists Press, Palo Alto, CA, USA.

[6] Fuchs, T., Haber, J., Seidel and H. P., 2004. MIMIC — A Language for Specifying Facial Animations. *Proc. of the 12th International Conf. in Central Europe on Computer Graphics, Visualization, and Computer Vision.* Plzen-Bory, Czech Republic, pp. 71-78.

[7] Guenter, B., Grimm, C., Wood, D., Malvar, H. and Pighin, F., 1998. Making Faces. ACM *Siggraph proc*. pp. 55-66.

[8] Hijazi, Y., 2005. *Integrating Speech Synthesis and Emotional Variations with Dynamic 3D Facial Expressions using Simulation of Anatomical Muscles*. MS Thesis, Kent State University, Kent, OH, USA, 149 pages.

[9] Hoppe, H., 1996. Progressive Meshes. *Proceedings of the 23rd annual Conf. on Computer graphics and interactive techniques*. New York, USA, pp. 99-108.

[10] Huang, Z., Eliens A. and Visser, C., 2003. XSTEP: An XML-based Markup Language for Embodied Agents. *Proc. of the 16th International Conf. on Computer Animation and Social Agents.* New Brunswick, USA, pp. 105-110.

[11] Kähler, K., Haber, J. and Seidel, H. P., 2003. Reanimating the Dead: Reconstruction of Expressive Faces from Skull Data, *ACM Transactions on Graphics*, Vol. 22, No. 3, pp. 554-561.

[12] Kim, J., Jordan, J. and Srinivasan, M. A. et al., 2004. Transatlantic Touch: A Study of Haptic Collaboration over Long Distance, *Presence: Teleoperators & Virtual Environments*, Vol. 13, No. 3, Pages 328-337.

[13] Magnenat-Thalmann, N., Primeau, N. E. and Thalmann, D., 1988. Abstract muscle actions procedures for human face animation. *Visual Computer*, Vol. 3, Number 5, pp. 290–297.

[14] Microsoft Speech and Speech SDK 5.1, Available [online]: http://www.microsoft.com/speech/

[15] Moubaraki, L. and Ohya, J., 1996. Realistic 3D Mouth Animation Using a Minimal Number of Parameters. *Proc. of the IEEE International Workshop on Robot and Human Communication*. pp. 201–206.

[16] Nakatsu, R., Tosa, N. and Ochi, T., 1998. Interactive Movie System with Multi-person Participation and anytime Interaction Capabilities. *Proc. 6th ACM Conf. on Multimedia: Technologies for interactive movies*. Bristol, UK, pp. 2-10.

[17] Noh, J. and Neumann, U., A Survey of Facial Modeling and Animation Techniques, http://graphics.usc.edu/cgit/pdf/papers/survey.pdf.

[18] Parke, F. I., 1972. Computer Generated Animation of Faces. *Proc. ACM annual conference.* Boston, USA, pp. 451-457.

[19] Parke, F. I. and Waters, K., 1996. Computer Facial Animation. Publisher: A. K. Peters Ltd., ISBN 1-56881-014-8.

[20] Pighin, F., Hecker, J., Lischinski, D., Szeliski, R. and Salesin, D. H., 1998. Synthesizing Realistic Facial Expressions from Photographs. *Proc. of the Int. Conf. on Computer Graphics and Interactive Techniques.* ACM Press, pp. 75-84.

[21] Platt, S. and Badler, N., 1981. Animating facial expression. *Computer Graphics*, Vol. 15, No. 3, pp. 245-252.

[22] Reeves, W. T., 1990. Simple and Complex facial animation: Case Studies. *17th International Conference on Computer Graphics and Interactive Technique*. Dallas, USA, pp. 88-106.

[23] Simoes, B. and Bansal, A. K., 2004. Interactive 3D Dynamic Object Based Movies. *Proceedings of the Fifth International Conference on Internet Computing*. Las Vegas, USA, Vol. II, pp. 708-714.

[24] Viad, M. L. and Yahia, H., 1992. Facial animation with wrinkles. Proceedings of the 3rd Eurographics Workshop on Animation and Simulation. Cambridge, UK.

[25] VoiceXML, http://www.voicexml.org/

[26] Virtual Reality Modeling Language(VRML) 2.0, ISO/IEC 14772, *http://www.web3d.org/x3d/specifications/vrml/ISO_IEC_14772-All/index.html.*

[27] Walczak, K. and Cellary, W., 2002. X-VRML – XML Based Modeling of Virtual Reality. *Proc. of the IEEE Symposium on Applications and the Internet*. Nara City, Japan, pp. 204-213.

[28] Waters, K., 1987. A Muscle Model for Animating Three-Dimensional Facial Expressions. *Computer Graphics*, Vol. 21, Issue 4, pp. 17-24.

[29] Waters, K. and Frisbie, J., 1995. A Coordinated Muscle Model for Speech Animation. *Graphics Interface*. pp. 163 – 170.

[30] Waters, K. and Levergood, T. M., 1993. DECface: An Automatic Lip- Synchronization Algorithm for Synthetic Faces. *Tech. Report*, CRL 93/4, DEC Cambridge Research Lab., USA

[31] Wu, Y., Kalra, P., Moccozet, L. and Thalmann, N. M., 1999. Simulating Wrinkles and Skin Aging. *The Visual Computer*, Vol. 15, No. 4, pp. 183-198

[32] Zhang, Y., Prakash, E. C. and Sung, E., 2004. Face alive. *Journal of Visual Languages and Computing*, Vol. 15, No. 2, pp. 125-160.