Talking Avatar for Web-based Interfaces

José Nunes, Luís Sá and Fernando Perdigão
Instituto de Telecomunicações
Coimbra, Portugal
{josenunes, luis, fp}@co.it.pt

Abstract— **In this paper we present an approach for creating interactive and speaking avatar models, based on standard face images. We have started from a 3D human face model that can be adjusted to a particular face. In order to adjust the 3D model from a 2D image, a new method with 2 steps is presented. First, a process based on Procrustes analysis is applied in order to find the best match for input key points, obtaining the rotation, translation and scale needed to best fit the model to the photo. Then, using the resulting model we refine the face mesh by applying a linear transform on each vertex. In terms of visual speech animation, we have considered a total of 15 different positions to accurately model the articulation of Portuguese language – the visemes. For normalization purposes, each viseme is defined from the generic neutral face. The animation process is visually represented with linear time interpolation, given a sequence of visemes and its instants of occurrence.**

Keywords: talking heads; speech animation;model adjustment;

I. INTRODUCTION

Human-computer interaction plays an increasingly important role on today's computer systems. The use of virtual animated characters on current digital support systems can greatly benefit user experience and interaction. These virtual characters, or simply avatars, can be applied on a wide range of applications for entertainment, personal communications, commerce, or education [1]. Along with the development of new web standards and technologies, today it is possible to deploy standard computer graphics applications on Internet environments, keeping good balance between visual quality and latency [2]. The solution we present in this work is primarily a system for avatar creation and animation, which can be deployed on web platforms. Although there are several approaches concerning model presentation, i.e. [1, 3] and model deformation, [4, 5, 6], this work presents a new approach for avatar creation, which is achieved only with a photo with no special requirements and no previous learning.

II. SYSTEM OVERVIEW

In this section we present a general view about the avatar framework and how it can be used in order to enhance web interface systems. As a distributed application in client-server architecture, there is a Graphical User Interface (GUI) that includes visual models, media and animation, and a server which provides a range of services. Thus, for applications using avatar, it is adopted the communication model shown on Fig. 1.

In this model, one or more clients can be connected simultaneously, accessing a webpage where the avatar is presented. This site is the GUI that will give users' access to visual information, such as models, images and their animation, to media which is basically audio playback, and to interaction, that is the set of interactive user controls, such as buttons and toolboxes. This application relies on services provided by a remote server. These services could be related in this case to audio synthesis, phonetic transcription and viseme conversion, semantics and adjustment algorithms.

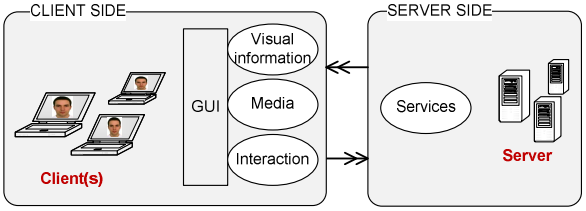

Figure 1. System Communication using Avatar

Apart from the main application, there is also a program to create users' own avatars from a face image. This is a client-server application, which has a GUI for choosing and adjusting points on a photo, and web services for model computation.

III. BASE MODEL

In order to represent all types of faces, regardless of gender, race or age, we have modeled a generic human head model, which resulted on a simplified representation of the face region, with always the same depth information. Topologically, this model consists on a polygon mesh with 152 connected vertices, forming 276 triangular faces. Since all human faces are nearly symmetric, first we have adjusted and simplified one side, and then the other side was obtained by mirroring. The mouth region, used for speech animation has 23 vertices. We have independent models for eye, tongue and teeth. The model is shown on Fig. 2 from 3 views.

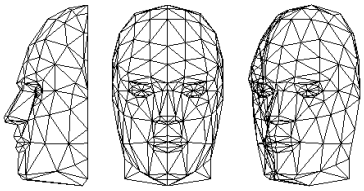

Figure 2. Base model polygonal mesh

This model will be textured from a photo. The ears and hair are not part of the mesh because they are not prominent in depth, so they tend to stay in the back of the model even while animating. Moreover, it is common that the ears appear partially or fully covered on many face images. Finally, ears and hair would be extremely difficult to adjust by automatic processes. The model's low complexity makes it suitable for internet scenarios, and generally for situations with no high-level processing. Each vertex and face is labeled so it is possible to assign regions of interest for animation and interaction. The processes of deformation and adjustment from a photo are described below.

IV. MODEL ADJUSTMENT FROM A PHOTO

A. Overview

In this section we present our approach to adjust and deform the base model in order to obtain the fitted mesh from a photo. This mesh adjustment is done with 2 different steps, which corresponds to global and local deformation [4]. We use 37 key points from critical regions like eyes, nose and mouth. These points must be identified manually (or automatically) in the photo and are used to deform the 3D model to best match the photo using both a global and a local adjustment.

B. Procrustes Analysis

Having the base model described on Section III we want to rotate and deform the model in such a way that its projection in the plane of the photo makes the best matching to the 2D key points on the photo. Mathematically this corresponds to a projection Procrustes problem [7]. The 3D (column) points $\mathbf{x}$ are projected into a 2D points $\mathbf{y}$ according to the linear transformation

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}, \tag{1}$$

where $\mathbf{b}$ is a translation and $\mathbf{A}$ is a 2x3 orthonormal projection matrix – with the restriction $\mathbf{A}\mathbf{A}^T = \mathbf{I}$. We can augment the $\mathbf{y}$ points with null z coordinate and use a 3x3 projection (idempotent) matrix $\mathbf{A}^T\mathbf{A}$ which projects 3D points according the original 3D axes. The unrestricted linear transformation problem has the following least squares solution:

$$\mathbf{A} = \mathbf{C}_{yx}\mathbf{C}_x^{-1} \;;\; \mathbf{b} = \boldsymbol{\mu}_y - \mathbf{A}\boldsymbol{\mu}_x, \tag{2}$$

where $\mathbf{C}_x$ is the covariance matrix of $\mathbf{x}$, $\mathbf{C}_{yx}$ the cross-covariance between $\mathbf{y}$ and $\mathbf{x}$ and $_x$ and $_y$ their mean vectors. We want to find a transformation matrix $\mathbf{A}$ in the form $\mathbf{A} = \mathbf{PSQ}$, where $\mathbf{P}$ is the 3D projection matrix, $\mathbf{S}$ is a scaling matrix (ideally diagonal), and $\mathbf{Q}$ is a rotation matrix. There is no close-form solution for this problem; however, an approximate solution can be taken, using the following SVD decomposition:

$$\mathbf{C}_{xy}\mathbf{C}_x^{-1} = \mathbf{U}\mathbf{L}\mathbf{V}^T = \mathbf{U}\mathbf{L}\mathbf{U}^T\mathbf{U}\mathbf{V}^T. \tag{3}$$

The matrix $\mathbf{Q} = \mathbf{U}\mathbf{V}^T$ is our rotation matrix. If det($\mathbf{Q}$) = -1, we can exchange the sign of the elements of the last row, in order to guarantee that it is a rotation matrix, without altering the solution. The scaling matrix will be $\mathbf{S} = \mathbf{U}\mathbf{L}\mathbf{U}^T$, which is symmetrical, non-negative defined, but not necessarily diagonal. There is a last problem to solve. Because matrix $\mathbf{S}$ has a last row and column with zeros, we need to estimate a scaling value for the z coordinate. This coordinate must shrink or grow according to the other two coordinates, so we choose to take it as the geometric mean of the main diagonal values of the matrix $\mathbf{S}$,

$$S_{33} = s_z = \sqrt{S_{11}S_{22}}, \tag{4}$$

and change $\mathbf{S}$ accordingly. We found that this scaling matrix is always close to a diagonal matrix.

Given the matrices $\mathbf{S}$ and $\mathbf{R}$ and the vector $\mathbf{b}$, we can apply them to all the model points, in order to get the first stage adjusted model. An example of an adjusted (and projected) model is presented on Fig. 3a), for a given photo. The key points are represented in green. We can see that the model was correctly scaled and rotated for the given photo. However, there is no precise match between key points and the projected model points because the solution is taken as a least squares problem. This justifies a second step of the method, described below.

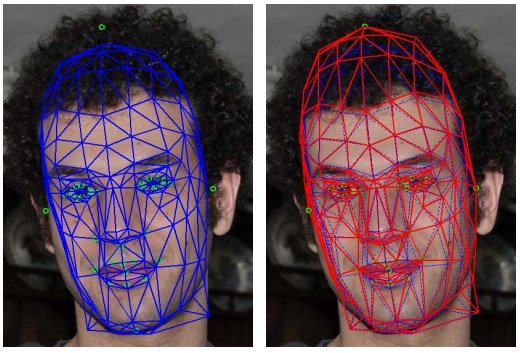

Figure 3. a) Model after Procrustes Analysis; b) Model after local deformation.

C. Local deformation

After the computation of global transformation, we need to adjust the model locally in order to accurately match the selected key points with the corresponding ones from the 3D model obtained after global transformation and projection. The first step consists on finding a 2D triangular mesh for each structure of key points – both for those placed on photo and the ones from 3D model. These meshes are computed from the Delaunay triangulation algorithm [8]. Our solution consists in a set of transformations that maps each model triangle into the corresponding photo key point triangle. All other model points are transformed in this way, being sufficient to find the triangle which the point belongs to or is interior to. This corresponds to a linear transform of the form $\mathbf{y}=\mathbf{T}_i\mathbf{x}+\mathbf{t}_i$ where $\mathbf{T}_i$ is the 2x2 matrix and $\mathbf{t}_i$ vector is a 2D point, given by

$$[\mathbf{T}_i \mid \mathbf{t}_i] = [\mathbf{y}_{1i} \quad \mathbf{y}_{2i} \quad \mathbf{y}_{3i}] \begin{bmatrix} \mathbf{x}_{1i} & \mathbf{x}_{2i} & \mathbf{x}_{3i} \\ 1 & 1 & 1 \end{bmatrix}^{-1}, \tag{5}$$

where $\mathbf{x}_{ji}$ and $\mathbf{y}_{ji}$ are the 3 vertices of the triangle i from model and photo, respectively. The model's 3D points are obtained simply by attaching its previous z coordinates.

Fig. 3b) shows the final model, in red, after local and global adjustment. The key points are shown in green. We see that the full adjusted red model passes on these points. The model shown on blue is the result from the first step of the adjustment algorithm, as already shown on Fig. 3a).

We believe that a good result is achieved after global and local adjustments. The avatar final quality depends greatly on key point placement. Fig. 4 shows an adjusted and full textured avatar model, with opened mouth and with eyes and model teeth.

Figure 4. Full adjusted avatar model with texture mapping

In order to keep the photo background on created avatar, we augment the resulting 3D mesh with additional vertices and corresponding faces. These new points are placed on the edges of the model so that when texturing, all pixel data is mapped.

V. ANIMATION WITH SPEECH

A. Overview

In this section we present one of the system's key features, which concerns the avatar's lip animation synchronized with speech. This issue is closely related to speech processing and synthesis. The phonetic information extracted from text or speech is necessary to prepare mouth animation and dynamics along time. The extracted phonemes are the basic units of speech and have visual equivalent counterparts, named visemes. On the other hand, visemes are unique face, mouth, teeth and tongue movements corresponding to voice phonemes [9]. To achieve smooth animation when synthesizing visual speech, various methods can be applied. Generally, these methods are based on data interpolation [9].

B. Visemes

In this section we present our approach for viseme mapping, modeling and normalization. According to [3], the relation between European Portuguese phones and their visual representation can be modeled with 15 correspondences. Using this information, the phone to viseme mapping is processed using a table. In order to produce smooth animation, the 15 visemes were manually adjusted for the base avatar model. Then, each viseme is transformed according to model adjustment, in order to fit all the positions to final avatar. Fig. 5 shows the base model adjusted to represent visemes 'p, b, m' and 'O' respectively.

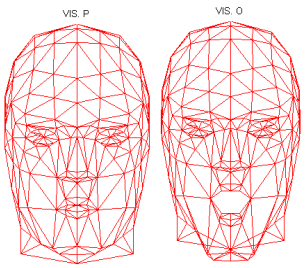

Figure 5. Visemes for phones "p, b, m" and "O"

To support viseme creation and modeling, several images and videos were analyzed in terms of mouth shape and dynamics. Such study also takes into account the visibility and positions for tongue and teeth. This visual information was extracted from different speakers, on various phonetic contexts. Some portions of those images are shown on Fig. 6.

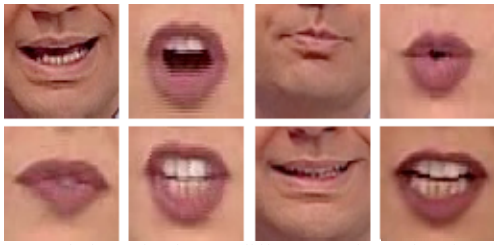

Figure 6. Speech dynamics for two different speakers taken from broadcast news in Portuguese

As mentioned above, we have considered a limited amount of points to model the visemes for the avatar. The collection of 23 points, shown on Fig. 7 is enough to assure accurate mouth animation. Note that, because of their proximity, some points on mouth region are not visible in the figure below.

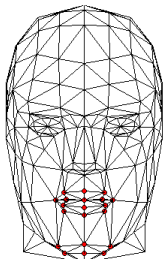

Figure 7. Viseme model points

Teeth and tongue play also an important role during speech. Its position relative to other mouth elements and its visibility are essential to display correct speech posture. Each viseme is defined as a data structure containing a unique identifier, its time of action, information regarding associated phones and any stress marks that may occur.

C. Animation

The animation process is mediated by a collection of visemes, which is defined in this context, as the data structure described above. The unique viseme identifiers and its temporal information, combined with animation algorithms

based on linear interpolation are enough for smooth speech articulation. The algorithm relies on a state computation basis, where previous and next visemes are considered to set the current positions. Thus, on each frame the current time is determined, the value between previous and next viseme is computed and then the forthcoming state is set. Fig. 8 presents the interpolation for the Portuguese word /Ola/ with a constant frame rate of 20 fps (frames per second). The horizontal axis represents the time, in milliseconds, and the vertical axis shows the Y coordinate value of a model point. The red diamonds are the Y values on each frame – the interpolated value between the visemes.

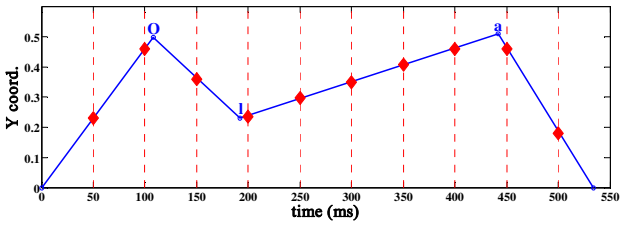

Figure 8. Interpolation for the Portuguese word /Ola/

We believe that more complex visemes may be introduced. Still, using linear interpolation with one position per viseme, the system shows fair results on speech articulation. We have seen that realism concerning speaking process is not dependent only upon viseme quantity and quality.

Apart from speech articulation, there are also other animation mechanisms to enhance avatar realism, which are essentially eye blinking and head or eye movements. Since there are also emotional and conversation signals that are important to realistic communication we already have a primary expression synthesizer that will be integrated on the speaking process.

VI. Implementation Issues

The solution is implemented with Microsoft Silverlight along with other external tools and libraries. Silverlight is a web application framework with support for media integration, multithreading and web services, which makes it useful for developing a wide range of applications [10]. The speech processing and synthesis is handled by Loquendo's TTS (Text-to-Speech) system [11]. To handle more complex data structures, like matrices, vectors, as well as their operations and transforms, we use Math Dot Net [12]. The Delaunay Triangulation is obtained by using planar subdivision routines from EmguCV [13]. In order to guarantee data conforming between client application and server, we have created an avatar description file in XML (Extended Markup Language), which has all the information needed to display and animate the model – geometric data, textures, transformation, constants, visemes and expressions. To assist the key point placement process, we use Stasm [14], which is a library for facial feature extraction, based on Active Shape Models.

VII. Conclusions and Future Work

We have proposed in this paper a new method to create talking avatar models, focusing on the adjustment methods needed, in order to produce realistic faces from one image. It is always a challenge to model and reproduce any human structure or behavior on computer systems. Yet, the applied adjustment methods have proven to be accurate on model deformation, when using good correspondences between model and image. We have described a method to solve this problem with only one photo and no depth information. Even using pictures presenting slightly rotated faces, our approach can track that rotation in order to best fit the final model.

We consider that in further developments, it would be desirable to enhance the current key point extraction system, through a fully automatic and reliable process, in order to have lesser user intervention in avatar creation. In the present system, the animation mechanism is only dependent upon TTS output. Therefore, we are currently working on speech processing, in order to extract phonetic information from a speech signal. With the extracted phones and its times, we would be able to prepare avatar speech animation with audio only.

Acknowledgments

J. Nunes thanks Instituto de Telecomunicações for research grant on project AVATAR and Inogate for supporting.

References

[1] Igor S. Pandzic, "Facial animation framework for the web and mobile platforms" in Proc. 7th Int. Conf. on 3D Web Technology, Tempe, Arizona, USA, 2002, pp. 27-34

[2] Rynson W. H. Lau *et al.*, "Emerging web graphics standards and technologies" in IEEE Comp. Graph. and App., 2003, pp. 66-75

[3] J. Neto, R. Cassaca, M. Viveiros, M. Mourão, "Design of a multimodal input interface for a dialogue system", in Proc. Int. Conf. Process. of Portuguese, 2006, Rio de Janeiro, Brasil, 2006

[4] A. Ansari and M. Adbel-Mottaleb, "3D Face Modelling using two orthogonal views and a generic face model" in Proc. Int. Conf. Multimedia and Expo, Baltimore, 2003, pp. 289-292

[5] K. Lin, F. Wang, J. Yao, C. Zhou, "Human Head Modeling Based on an Improved Generic Model", vol. 4, pp.300-304, in Proc. 5th Int. Conf. Fuzzy Systems and Knowledge Discovery, 2008

[6] K. Zhang, Z. Huay, T. Chua, A Framework to custumize a face model for reusing animation, in Proc. Comput. Graph. Int. 2003, Tokyo, Japan

[7] J. C. Gower and G. B. Dijksterhuis, *Procrustes Problems*, Oxford University Press, 2004 (ISBN 0198510581)

[8] D. T. Lee and B. J. Schachter, "Two algorithms for constructing a Delaunay triangulation." Int. J. Comp. Inform. Sci. 9, 219-242, 1980

[9] F.I. Parke and K. Waters, "Speech Synchronized Animation" in Computer Facial Animation, 2nd ed, Wellesley MA, AK Peters, 2008, pp. 296-300

[10] Microsoft, "Silverlight Overview,", Internet: http://www.microsoft.com/silverlight/what-is-silverlight, [Nov. 11, 2010]

[11] P. Baggia, S. Mosso, "Speech technologies and multimodality: The solution for new advanced services", white-paper, April 2005

[12] Christoph Rüegg, "Math.NET Project", Internet: http://www.mathdotnet.com, Sep. 6, 2010, [Nov. 11, 2010]

[13] EmguCV Project, "An Intel OpenCV .NET Wrapper", Internet: http://www.emgu.com/wiki/index.php/Main_Page, April 15, 2010, [Nov. 11, 2010]

[14] S. Milborrow, F. Nicolls, Locating facial features with an Extended Active Shape Model, European Conf. Comput. Vision, 2008