LUCIA-WebGL: A Web Based Italian MPEG-4 Talking Head

G.Riccardo Leone, Piero Cosi

Institute of Cognitive Sciences and Technologies - National Research Council of Italy
`riccardo.leone@pd.istc.cnr.it, piero.cosi@pd.istc.cnr.it`

Abstract

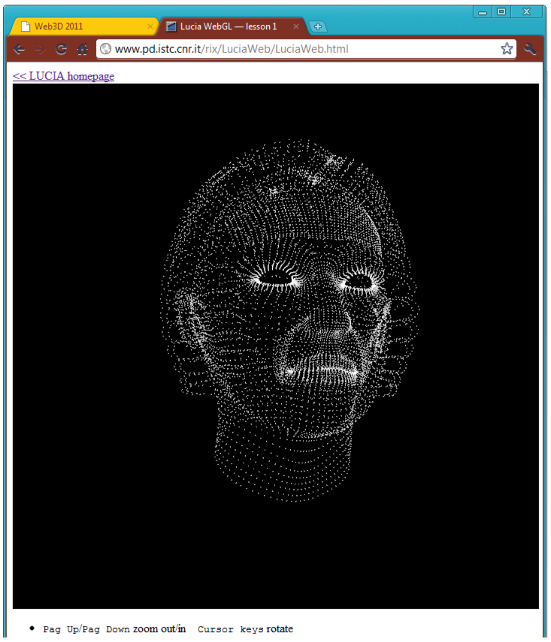
In this work we present the reviewing of the activities focused on the development of the WebGL software version of LUCIA talking head, an open source facial animation framework developed at ISTC-CNR of Padua. LUCIA works on standard MPEG-4 Facial Animation Parameters and speaks with the Italian version of FESTIVAL TTS. LUCIA is totally based on true real human data collected by the use of ELITE, a fully automatic movement analyzer for 3D kinematics data acquisition. These informations are used to create lips articulatory model and to drive directly the talking face, generating human facial movements. We are exploiting the use of LUCIA WebGL as a virtual guide in the Wikimemo.it project: The portal of Italian Language and Culture. The easy integration of this technology in websites offers promising future uses for the WebGL Avatars: on-line personal assistant, storyteller for web-books, digital tutor for hearing impaired are only few examples.

Index Terms: WebGL, talking head, facial animation, mpeg4, 3D avatar, virtual agent, TTS, LUCIA, FESTIVAL

1. Introduction

Computers are becoming an integral part of all activities in our daily lives and the simply natural interaction with them is one of the most vibrant research in the last decade. Life-like characters [1] are embodied agents living on the screens of computational devices that invite us to communicate with them in familiar, expressive/emotive multi-modal ways. Face to face communication is the main element of human-human interaction because both acoustic and visual signal simultaneously convey linguistic, extra linguistic and paralinguistic informations. This is the reason why a realistic audio/visual synthesis is very important for virtual agents. This is a research topic since the early 70's and many different principles, models and animations have been proposed over years [2]. In the late 90's a specification for efficient coding of shape and animation of human face was included in the MPEG-4 international standard [3]. The focus was extended from traditional audio and video coding to other multimedia context including images, text, graphics, 3D scenes, animation and synthetic audio. Concerning Facial Animation MPEG-4 standard defines the shape of the model (FDP) and a set of actions (FAP); the animation is obtained by specifying a stream of numbers that is for each frame the values of the Facial Animation Parameters. Many implementations of this standard were born in the last decade [4] but they often remain stand alone applications built for research purpose. The recent introduction of WebGL [5], which is 3D graphics in web browsers, opens the possibility to bring all these applications to the home computers of a very large number of persons and to burst this natural way of interaction with the machines. WebGL extends the capability of the JavaScript programming language to generate interactive 3D graphics within any compatible web browser. The WebGL Working group (including Mozilla, Google, Apple and Opera) started in early 2009;two years after, on February 2011, they released the version 1.0 of WebGL. On May 2011 Google Chrome and Mozilla Firefox support WebGL and so will be in short time also Apple Safari and Opera browsers. WebGL is based on OpenGL Embedded System 2.0, the Graphic Library developed for mobile devices. This means that it does not have all the functionalities of the latest OpenGL for desktop, but on the other hand it is more easy that WebGL websites will run (in the near future) on smart-phone and other mobile devices. This is a real revolution because it brings the power of 3D graphics directly into web-browser without installing any plug-ins or customized and maybe dangerous software. The easy integration of this technology in any website offers promising future uses for WebGL Avatars: on-line personal assistant, storyteller for web-book, digital tutor for hearing impaired are only few examples.

Figure 1: Lucia talking head model vertices rendered in Chrome web-browser


Web3D 2011
Lucia WebGL — lesson 1
www.pd.istc.cnr.it/rix/LuciaWeb/LuciaWeb.html
<< LUCIA homepage
• Pag Up/Pag Down zoom out/in   Cursor keys rotate

2. The WebGL Client

Lucia-WebGL follows the common client-server paradigm. We have a client (a web browser) that connects to the server opening a web-page. This server answers with an HTML5 web-

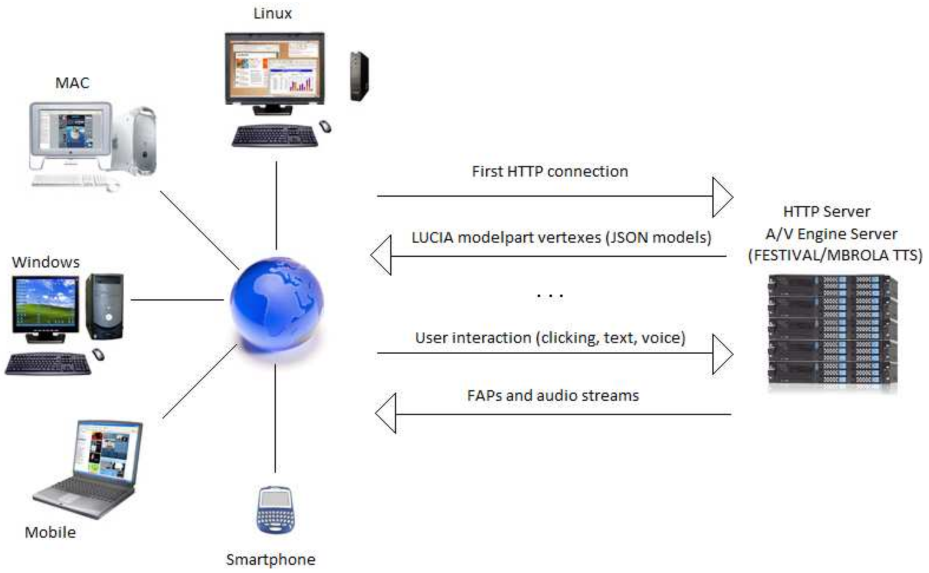


Figure 2: The new client-server architecture: WebGL allows any system, even smart-phone and P.D.A, to interact with LUCIA via standard web browsers. At the beginning of the connection the model-parts of Lucia are fetched from the server in the JSON format. After that every communication involves only FAPs and audio streams with a very low bandwidth consumption

page where there is a new element, the Canvas 3D, which is the place where all our 3D graphics lives. At the beginning of the connection the model parts data are sent on the Internet using the lightweight data-interchange format JSON (JavaScript Object Notation) [6]. This is the only moment where we can wait for a while because of the amount of the data to transmit. The main 3D model is composed by seven fundamental elements: the skin, the inner articulators (the tongue, the teeth and the palate) and the facial elements (the two eyes and the hair). LUCIA is a textured young female 3D face model built with 25423 polygons: 14116 belong to the skin, 4616 to the hair, 2688x2 to the eyes, 236 to the tongue and 1029 to the teeth respectively. The textures used for the model are shown in fig. 3 Once the 3D model has been loaded and rendered in the web-pages in its neutral position all communications between client and server are very low bandwidth consumption. This depends directly on MPEG-4 standard and on LUCIA animation mechanism. LUCIA emulates the functionalities of the mimic muscles, by the use of specific displacement functions and of their following action on the skin of the face. The activation of such functions is determined by specific parameters that encode small muscular actions acting on the face, and these actions can be modified in time in order to generate the wished animation. Such parameters, in MPEG-4, take the name of Facial Animation Parameters and their role is fundamental for achieving a natural movement. Moreover, the muscular action is made explicit by means of the deformation of a polygonal reticule built around some particular key points called Facial Definition Parameters (FDP) that correspond to the junction on the skin of the mimic muscles. FDPs define the shape of the model while FAPs (Facial Animation Parameters), define the facial actions and, given the shape of the model, the animation is obtained by specifying the FAP-stream that is for each frame the values of FAPs. In a FAP-stream, each frame has two lines of parameters. In the first line the activation of a particular marker is indicated (0, 1) while in the second, the target values, in terms of differences from the previous ones, are stored(fig. 5).

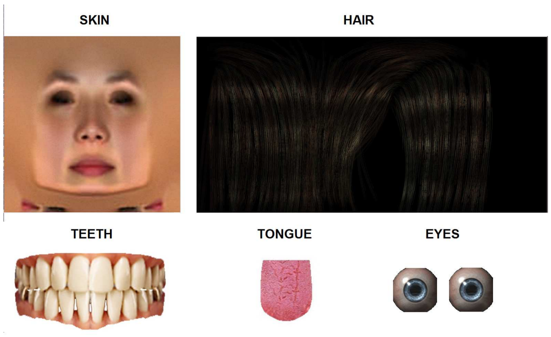


Figure 3: The textures used for LUCIA model

Moving only the FDPs is not sufficient to smoothly move the whole 3D model, thus, each ”feature point” is related to a particular influence zone constituted by an ellipses that represents a zone of the reticule where the movement of the vertexes is strictly connected. We consider two zone in the influence area: the points of area A that are affected by muscular contraction will be deformed by the muscular displacement function, while the points of area B (area of the bulge / furrow) will be moved outward to simulate the skin accumulation and bulging (fig. 4).

Finally, after having established the relationship for the whole set of FDPs and the whole set of vertexes, all the points of the 3D model can be simultaneously moved with a graded strength following a raised-cosine function rule associated to each FDP. Each feature point follows MPEG4 specications where a FAP corresponds to a minimal facial action. When a FAP is activated (i.e. when its intensity is not null) the feature point on which the FAP acts is moved in the direction signaled by the FAP it-self (up, down, left, right, etc). Using the pseudo-muscular approach, the facial models points within the region of this particular feature point get deformed. A facial expression

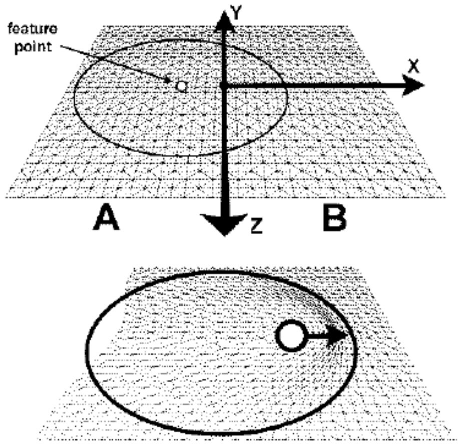
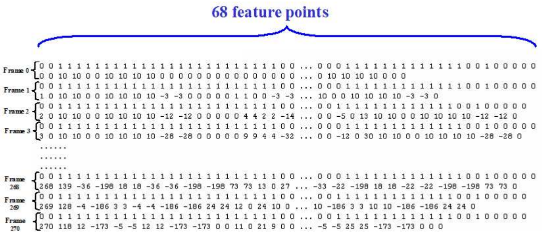


Figure 4: The skin deformation in the area of influence is achieve considering two zone: A (muscular traction) and B (accumulation)

is characterized not only by the muscular contraction that gives rise to it, but also by an intensity and a duration. The intensity factor is rendered by specifying an intensity for every FAP. The temporal factor is modeled by three parameters: onset, apex and offset [7]. Only the reticule of polygons corresponding to the skin is directly driven by the pseudo-muscles and it constitutes a continuous and unitary element, while the other anatomical components move themselves independently and in a rigid way, following translations and rotations (for example the eyes rotate around their center). According to this strategy the polygons are distributed in such a way that the resulting visual effect is quite smooth with no rigid jumps over the entire 3D model. The synchronization of the lips movements with the audio is achieved due to the frame number information in the FAPs stream. The FAPs stream and the audio stream are the necessary information to animate the MPEG-4 based synthetic talking face (see fig. 2). These streams are generated on the server side by the Audio Video Engine.


68 feature points

```
Frame 0   0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 ... 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0
          0 0 10 10 0 0 10 10 10 10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ... 0 10 10 10 10 0 0 0
Frame 1   0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 ... 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0
          1 0 10 10 0 0 10 10 10 10 -3 -3 0 0 0 0 0 1 1 0 0 -3 -3 ... 10 0 0 10 10 10 10 -3 -3 0
Frame 2   0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 ... 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0
          2 0 10 10 0 0 10 10 10 10 -12 -12 0 0 0 0 0 4 4 2 2 -14 ... 0 0 -5 0 13 10 10 0 0 10 10 10 10 -12 -12 0
Frame 3   0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 ... 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0
          3 0 10 10 0 0 10 10 10 10 -28 -28 0 0 0 0 0 9 9 4 4 -32 ... 0 0 -12 0 30 10 10 0 0 10 10 10 10 -28 -28 0
          ......
          ......
          ......
Frame     0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 ... 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0
 268      268 139 -36 -198 18 18 -36 -36 -198 -198 73 73 13 0 27 ... -33 -22 -198 18 18 -22 -22 -198 -198 73 73 0
Frame     0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 ... 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0
 269      269 128 -4 -186 3 3 -4 -4 -186 -186 24 24 12 0 24 10 0 ... 10 -186 3 3 10 10 -186 -186 24 24 0
Frame     0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 ... 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0
 270      270 118 12 -173 -5 -5 12 12 -173 -173 0 0 11 0 21 9 0 0 ... -5 -5 25 25 -173 -173 0 0 0
```

Figure 5: The FAPs stream is the necessary information to animate an MPEG-4 based synthetic face. It is a very low bandwidth transmission very good also with a slow connection

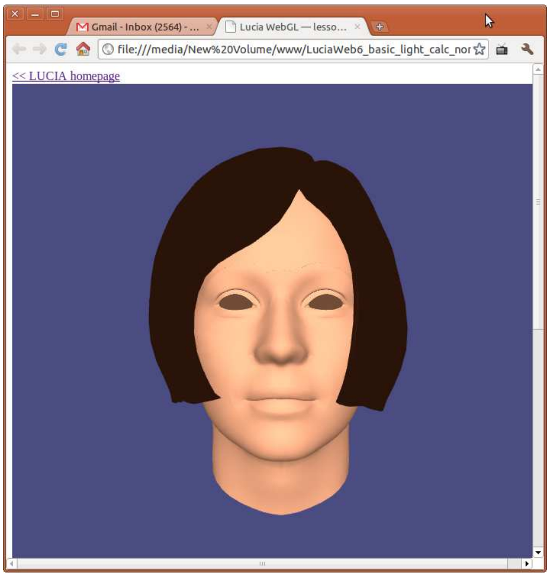
3. Audio Video Engine Server

Audio Video speech synthesis, that is the automatic generation of voice and facial animation parameters from arbitrary text, is based on parametric descriptions of both the acoustic and visual speech modalities. The acoustic speech synthesis uses an Italian version of the FESTIVAL di-phone TTS synthesizer [8] modified with emotive/expressive capabilities: the APML/VSML mark up language [11] for behavior specification permits to specify how to markup the verbal part of a dialog move so as to add to it the ”meanings” that the graphical and the speech generation components of an animated agent need to produce the required expressions. For the visual speech synthesis a data-driven procedure was utilized: visual data are physically extracted by an automatic opto-tracking movement analyzer for 3D kinematics data acquisition called ELITE [12]. ELITE provides 3D coordinate reconstruction starting from 2D perspective projections by means of a stereophotogrammetric procedure which allows a free positioning of the TV cameras. The 3D data coordinates of reflecting markers positioned on the model subject face are then used to create lips articulatory model. All the movements of the markers are recorded and collected, together with their velocity and acceleration, simultaneously with the co-produced speech which is usually segmented and analyzed by means of PRAAT [13], that computes also intensity, duration, spectrograms, formants, pitch synchronous F0, and various Voice Quality (VQ) parameters that are quite significant in characterizing emotive/expressive speech [14]. In order to simplify and automates many of the operation needed for building-up the 3D avatar from the motion-captured data we developed INTERFACE [15], an integrated software designed and implemented in Matlab. To generate realistic facial animation is necessary to reproduce the contextual variability due to the reciprocal influence of articulatory movements for the production of following phonemes. This phenomenon, defined co-articulation [16], is extremely complex and difficult to model. A variety of co-articulation strategies are possible and even different strategies may be needed for different languages [17]. A modified version of the Cohen-Massaro co-articulation model [9] has been adopted for LUCIA [18] and a semi-automatic minimization technique, working on the real cinematic data, was used for training the dynamic characteristics of the model, in order to be more accurate in reproducing the true human lip movements. The modified co-articulatory model is able to reproduce quite precisely the true cinematic movements of the articulatory parameters. The mean error between real and simulated trajectories for the whole set of parameters is, in fact, lower than 0.3 mm.

4. Conclusions and future work

LUCIA-WebGL is an MPEG-4 standard FAPs driven facial animation talking head implementing a decoder compatible with the ”Predictable Facial Animation Object Profile” (very early result in fig. 6). It has an high quality 3D model and a fine co-articulation model, which is automatically trained by real data, used to animate the face. The modified co-articulatory model is able to reproduce quite precisely the true cinematic movements of the articulatory parameters. The WebGL client-server architecture separates the visualization/interaction process (on the WebGL client) from the generation of the audio/visual streams necessary for the animation of the talking head (on the server side). These streams are very low bit-rate and can function very well also with slow connections. The next step after the basic communication process will be the implementation of the emotive/expressive functionalities based on APML/VSML markup language. Future developement will include the personalization of the model using as textures some photos of a real face taken from different views. The first use of LUCIA WebGL will be a virtual guide in the Wikimemo.it project: The portal of Italian Language and Culture. However the easy integration of this technology in common websites allows many other roles for LUCIA: we want to test it as a storyteller for web-books and a digital tutor for the hearing impaired persons.

Figure 6: Lucia-Webgl talking head in the neutral position


Gmail - Inbox (2564) - ...
Lucia WebGL — lesso...
file:///media/New%20Volume/www/LuciaWeb6_basic_light_calc_nor
<< LUCIA homepage

5. Acknowledgements

This work has been sponsored by WIKIMEMO.IT (The Portal of Italian Language and Culture, FIRB Project, RBNE078K93, Italian Ministry of University and Scientific Research)

6. References

[1] H. Prendinger and M. Ishizuka, Eds., *Life-Like Characters: Tools, Affective Functions, and Applications*. Berlin: Springer, 2004.

[2] F. I. Parke and K. Waters, *Computer facial animation*. Natick, MA, USA: A. K. Peters, Ltd., 1996.

[3] MPEG-4, http://mpeg.chiariglione.org/standards/mpeg-4/mpeg-4.htm.

[4] I. S. Pandzic and R. Forchheimer, Eds., *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. New York, NY, USA: John Wiley & Sons, Inc., 2003.

[5] WebGL, http://www.khronos.org/webgl/.

[6] JSON, http://www.json.org/.

[7] P. Ekman and W. Friesen, "Facial action coding system," *Consulting Psychologist*, 1978.

[8] P. Cosi, F. Tesser, R. Gretter, and C. Avesani, "Festival speaks italian!" in *Proceedings of Eurospeech 2001*. Aalborg, Denmark: Eurospeech, 2001, pp. 509–512.

[9] P. Cosi and G. Perin, "Labial coarticulation modeling for realistic facial animation," in *Proceedings of ICMI 2002*. Pittsburgh, USA: ICMI, 2002, pp. 505–510.

[10] LUCIA, http://www2.pd.istc.cnr.it/LUCIA/.

[11] B. D. Carolis, C. Pelachaud, I. Poggi, and M. Steedman, "Apml, a mark-up language for believable behavior generation," *Life-Like Characters*, pp. 65–85, 2004.

[12] G. Ferrigno and A. Pedotti, "Elite: A digital dedicated hardware system for movement analysis via real-time tv signal processing," *IEEE Transactions on Biomedical Engineering*, pp. 943–950, 1985.

[13] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, pp. 341–345, 1996.

[14] C. Drioli, P. Cosi, F. Tesser, and G. Tisato, "Emotions and voice quality: Experiments with sinusoidal modeling," in *Proceedings of Voqual 2003*. Geneva, Switzerland: ISCA, 2003, pp. 127–132.

[15] G. Tisato, C. Drioli, P. Cosi, and F. Tesser, "Interface: a new tool for building emotive/expressive talking heads," in *Proceedings of INTERSPEECH 2005*. Lisbon, Portugal: INTERSPEECH, 2005, pp. 781–784.

[16] E. Farnetani and D. Recasens, "Coarticulation models in recent speech production theories," *Coarticulation in Speech Production*, 1999.

[17] R. Bladon and A. Al-Bamerni, "Coarticulation resistance in english," *Phonetics*, pp. 135–150, 1976.

[18] P. Cosi, A. Fusaro, and G. Tisato, "Lucia a new italian talking-head based on a modified cohen-massaros labial coarticulation model," in *Proceedings of Eurospeech 2003*. Geneva, Switzerland: Eurospeech, 2003, pp. 2269–2272.

[19] D. Massaro, *Perceiving Talking Faces: from Speech Perception to a Behavioral Principle*. Cambridge, USA: MIT Press, 1997.

[20] E. Costantini, F. Pianesi, and P. Cosi, "Evaluation of synthetic faces: Human recognition of emotional facial displays," in *Tutorial and Research Workshop Affective Dialogue Systems*, Kloster Irsee, Germany, 2004, pp. 276–287.