TEXT TO VISUAL SYNTHESIS WITH APPEARANCE MODELS

Javier Melenchón, Fernando de la Torre, Ignasi Iriondo,
Francesc Alías, Elisa Martínez and Lluis Vicent

La Salle School of Engineering, Ramon Llull University
Pg. Bonanova 8, 08022 Barcelona, Spain
{jmelen, ftorre, iriondo, falias, elisa, vicent}@salleURL.edu

ABSTRACT

This paper presents a new method named text to visual synthesis with appearance models (TEVISAM) for generating videorealistic talking heads. In a first step, the system learns a person-specific facial appearance model (PSFAM) automatically. PSFAM allows modeling all facial components (e.g. eyes, mouth, etc) independently and it will be used to animate the face from the input text dynamically. As reported by other researches, one of the key aspects in visual synthesis is the coarticulation effect. To solve such a problem, we introduce a new interpolation method in the high dimensional space of appearance allowing to create photorealistic and videorealistic avatars. In this work, preliminary experiments synthesizing virtual avatars from text are reported. Summarizing, in this paper we introduce three novelties: first, we make use of color PSFAM to animate virtual avatars; second, we introduce a non-linear high dimensional interpolation to achieve videorealistic animations; finally, this method allows to generate new expressions modeling the different facial elements.

1. INTRODUCTION

The interaction with automatic systems is involving more tasks that can be done with computers. In order to enhance this interactivity, talking heads can be useful to communicate in a more natural way.

In the last years, many researchers have been working in the representation and the personalization of talking heads based on 2D models ([1], [2], [3], [4], [5]). In most of these previous works, the generation of the facial model involved some tedious manual tasks.

In the context of the animation from text, recently, a semi-automatic method has been proposed to avoid the manual cropping (the masks are specified by hand) [6]. This method is based on multidimensional morphable models (MMM). Also, De la Torre and Black [7] have proposed a semi-automatic learning process for
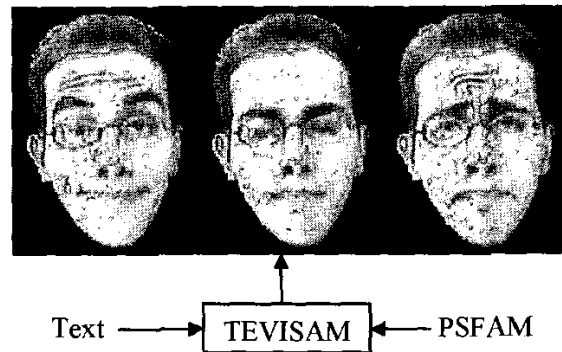

Figure 1. Inputs and outputs of TEVISAM

learning person-specific facial appearance models similar in spirit to MMM. Unlike morphable models, PSFAM allows better modeling of the facial features such as mouth, tongue, etc, since does not rely on noisy computation of optical flow. Also, in PSFAM the registration is done with respect to the optimal eigenspace constructing a better model.

Figure 2. Specification of the masks in the first frame.

On the other hand, one important aspect to create talking heads is the dynamic simulation. The dynamics generation covers the coarticulation effects [8] and it depends on the representation of the appearance. Representations based on a finite set of samples used in some works [1][3], rely on morphing techniques (which may result in jerky movements) for creating novel images. On the other hand, techniques based on a subspace of appearance [2][5] use a parametric representation. The generation of a novel image stream requires the specification of the proper trajectory through the subspace of appearance; the trajectory must be contained in the subspace following the actual dynamics of the face [9].

In this paper we use color PSFAM learned semi-automatically to create talking heads in Catalan from text, so just the masks have to be specified by the user (figure 2). Also, a non-linear high dimensional interpolation to achieve videorealistic animations is developed.

2. CONSTRUCTING PSFAM

Let $I = [i_1 \ \cdots \ i_m \ \cdots \ i_M]$ be a sequence of M vectorized images, where each i_m is an image taken at time m. Also, let $\Pi = [\pi^1 \ \cdots \ \pi^l \ \cdots \ \pi^L]$ be the set of L non-overlapping masks where each π^l is a vectorized binary image that specifies the spatial domain of the l-th facial element. Notice that all vectors of this paper are defined as column vectors.

The model of appearance is generated applying PSFAM [7], which provides an independent representation of every facial element. An image will be modeled as follows:

$$\hat{I}_m = \sum_{l=1}^{L} \left(\pi^l \circ B^l c_m^l \right) \left(f(x, a_m) \right) \tag{1}$$

where l is the index of the l-th facial element. $B^l = [b_1^l \ \cdots \ b_p^l \ \cdots \ b_{P_l}^l]$ is the orthogonal basis of appearance; c_m^l are the P_l coefficients whose linear combination of B^l reconstructs the appearance of the corresponding facial element; thus $\left(\pi^l \circ B^l c_m^l \right)$ represents the image of the facial element located in the spatial domain specified by the corresponding mask π^l; $f(x, a_m)$ is the affine model applied to represent the movement and the face alignment, where a_m are the 6 motion parameters and x are the Cartesian coordinates of the pixels.

The appearance bases B^l and the motion parameters are learnt simultaneously by an automatic process. A parametric motion model (coefficients a_m) is used to register the face. Before detailing the algorithm, some descriptions need to be specified:

- $A = [a_1 \ \cdots \ a_m \ \cdots \ a_M]$: affine parameters that align every image of the training set.
- $B = \{ B^1, \ \cdots, \ B^l, \ ..., \ B^L \}$: set of appearance bases. Each orthogonal basis can be defined as the set of eigenvectors which contains a determined value of energy.

- $C = \{ C^1, \ \cdots, \ C^l, \ ..., \ C^L \}$: the projection parameters. Let $C^l = [c_1^l \ \cdots \ c_m^l \ \cdots \ c_M^l]$ be the projection parameters c_m^l of the m-th image.

The appearance model is constructed by means of the following algorithm based on PSFAM

1. Manual initialization of the masks in the first image.
2. Registration of all images using only the first image obtaining A.
3. B are constructed for every facial element.
4. For each iteration
   4.1. Registration of all images using the previously learnt B obtaining the image projections C over the bases B and the affine parameters A.
   4.2. New appearances B are constructed for every facial element. This is not executed if we are in the last loop step.
5. The alignment, appearance bases and projections of the images are ready to be used at synthesis time.

In this approach, some details have been changed from the original PSFAM [7]. The first difference is in the initial iteration, where only the first image is used to register the full sequence without considering the initial bases B. This change speeds up the learning process. Also, unlike PSFAM, we do not use a stochastic initialization (e.g. genetic algorithms) because we assume that the motion of the user is not big enough. The energy of the eigenvectors of B is increased during the execution of the algorithm in order to gradually improve the description of the appearance model.

2.1 Adding color

The appearance is represented in color, but it can also be modeled in grey level. Using color in the tracking phase of the registration process achieves more accurate results at higher computational cost requirements. In this work, the color space RGB has been chosen, and we are working to validate the effectiveness of other color spaces in our model.

3. NON-LINEAR HIGH DIMENSIONAL INTERPOLATION

One key aspect for videorealistic animation is the coarticulation effect (that is the transition between phonemes)[8]. If this transition is not smooth enough the animation is not going to be credible. In order to improve this aspect, in this section, we will introduce a new non-

linear high dimensional interpolation in the appearance space for creating visually smooth trajectories.

Let $C^K = \begin{bmatrix} c_1^K & \cdots & c_m^K & \cdots & c_M^K \end{bmatrix}$ be the coefficients to reconstruct the mouth in the M images of the training sequence, recall that K indicates the facial element corresponding to the mouth.

A graph is an ordered pair $G = (V,E)$, where V is a finite, non-empty set of objects called vertices, and E is a (possibly empty) set of unordered pairs of distinct vertices i.e., 2-subsets of V called edges. Let G^K be a graph, with nodes all the set of coefficients from the training set C^K and links a set of weights defined by the Mahalanobis distance between coefficients. In order to compute the distance between coefficients we make use of the eigenspace computed in the training set. That is, let $I^K = B^K \Sigma^K \left(V^K\right)^T$ be the factorization of the set of aligned images of the mouth. The singular values (diagonal elements of Σ^K) are the square root of the eigenvalues, and taking into account that B and V are orthogonal matrices it is easy to show that the Mahalanobis distance between two coefficients is given by $G^K(i,j) = \left(c_i^K\right)^T \left(\Sigma^K\right)^{-2} c_j^K$. All necessary information to create visual coarticulation is located in G^K and it is used in section 4.

A correspondence between phonemes and visemes $v^K : phoneme(s_j) \rightarrow viseme(c_j^K)$ could be obtained automatically by means of a segmentation system. Notice that only about twenty correspondences (the same number as phonemes) are used in phoneme identification. Additional correspondences v^j could be defined in order to generate other facial expressions which take part in communication.

4. SYNTHESIS

Any utterance can be synthesized as a sort set of J points $\Gamma^K = \{\gamma_1^K, \ldots, \gamma_j^K, \ldots, \gamma_J^K\}$ expressed in the basis B^K. These points must belong to the previously learnt appearance subspace; otherwise, unnatural images may appear. In order to synthesize real visemes (trained or novel ones), the points γ_j^K have to describe a trajectory of minimum distance that will pass by some of the c_m^K points of the graf G^K using the shortest path algorithm [10]. This fact means that some c_m^K may be equal to some γ_j^K (see figure 3).
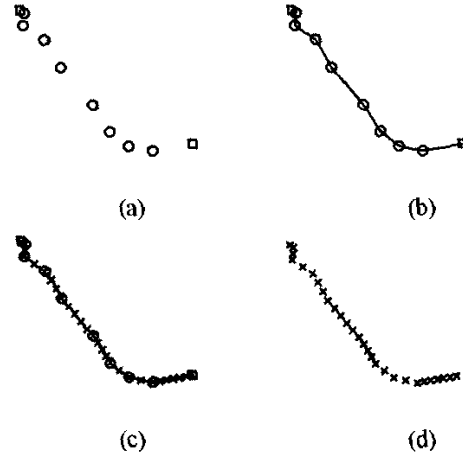

(a)   (b)

(c)   (d)

Figure 3. Visual coarticulation generation of the utterance /ðo/. (a) c_m^K points in the coarticulation trajectory, where $c_1^K = t_1^K = v^K(/\delta/)$, $c_R^K = t_2^K = v^K(/o/)$ and the remainder points are obtained by means of taking the dijkstra algorithm over the pair of nodes $\left(t_1^K, t_2^K\right)$. (b) Coarticulation trajectory $\widetilde{F_1}$. (c) Sampling of trajectory $\widetilde{F_1}$. (d) Points γ_j^K of the sampled trajectory $\widetilde{F_1}$; note that $\gamma_j^K = c_m^K$ for some j,m.

Given a stream $S = \{s_1, \ldots, s_q, \ldots, s_Q\}$ corresponding to a phoneme of an utterance U, let T^K be the set of Q points respect to B^K when the correspondence v^K is taken over S_j:

$$T^K = \{v^K(s_1), \ldots, v^K(s_q), \ldots, v^K(s_Q)\} =$$
$$= \{t_1^K, \ldots, t_q^K, \ldots, t_Q^K\} \quad (2)$$

The set T^K is the sampling of a trajectory F through the appearance subspace that represents the utterance U, but has no information about coarticulation. Visual coarticulation is modeled as the points that belongs to F but are not specified in T^K; these points will be obtained through the sampling of an approximation of F. Let $\widetilde{F} = \{\widetilde{F_1}, \ldots, \widetilde{F_q}, \ldots, \widetilde{F_{Q-1}}\}$ be an approximation of F that can be generated using the shortest path algorithm [10] on the graf G^K over every pair $\left(t_q^K, t_{q+1}^K\right)$ to obtain each segment $\widetilde{F_q}$. The final set Γ^K corresponding to utterance U will be the sampling of $\widetilde{F}$ (see figure 3). As a result, the trajectory $\widetilde{F_q}$ taken from t_q^K to t_{q+1}^K can be seen as an approximation of the geodesic trajectory between these two points; so, the visual coarticulation effect is more accurate than the

resulting linear interpolation between t_q^K and t_{q+1}^K (figure 3). A *training sequence* of at least 2 min. at 10 frames/sec is needed to have enough density of c_m^K in order to generate accurate visual coarticulation. Another *approximation of this geodesic trajectory can be seen in* [9].

Movements can be achieved with an affine model and cubic b-splines in a similar way than [5] *Translation, scaling and some rotations can be* synthesized with this model.

5. RESULTS

Some results of analysis and synthesis of talking heads using our model are shown. The registration process and its synthesized results can be seen in the figure 4. Novel utterances not observed in the training sequence are revealed in figure 5. As can be seen, the synthesized sequences achieve to show fine details as mouth, teeth and tongue. Generation of transition between facial expressions can be seen in figure 1.
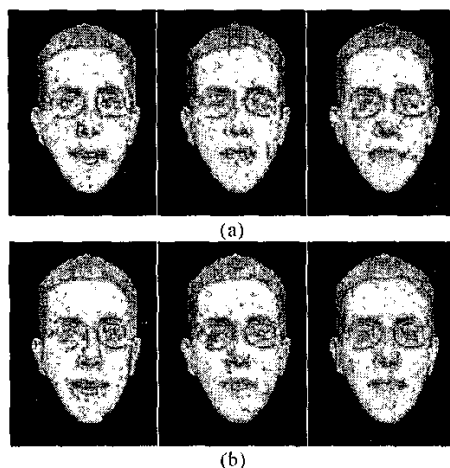

(a)

(b)

Figure 4. (a) Three registered frames. (b) The same synthesized frames

Figure 5. Utterance /ðo/ previously unseen

6. CONCLUDING REMARKS

This paper has introduced a new approach for modeling talking heads, which simplify the training process. The specification of the masks in the first image is the only manual part of the algorithm. Also, an approximation of geodesic distance in eigenspaces is proposed. However, there is still a substantial amount of work in the automation of the masks, the flexibility of the cheek, the synthesis of 3D rotations and the behavior of different color spaces.

7. ACKNOWLEDGEMENTS

8. REFERENCES

[1] C. Bregler, M. Covell, M. Slaney, "Video Rewrite: Driving Visual Speech with Audio", *Proc. SIGGRAPH '97*, pp. 353-360.
[2] T. Ezzat, T. Poggio, "Facial Analysis and Synthesis Using Image-Based Models", *Proc 2nd Int. Conf. On Automatic Face and Gesture Recognition*, IEEE CS Press, 1996, pp. 116-121.
[3] E. Cossatto, H. Graf, "Sample-Based Synthesis of Photorealistic Talking Heads", *Proceedings of Computer Animation '98*, pp. 103-110, Philadelphia, Pennsylvania, 1998.
[4] J. Noh and U. Neumann. "Talking Faces". *IEEE International Conference on Multimedia and Expo (II) 2000*, pp. 627-630, New York, USA.
[5] J. Melenchón, F. Alias, I. Iriondo, "PREVIS: a Person-Specific Realistic Virtual Speaker", *IEEE International Conference on Multimedia and Expo (II) 2002*, Lausanne, Switzerland.
[6] T.Ezzat, G. Geiger, T. Poggio, "Trainable Videorealistic Speech Animation", *Proc. of ACM Siggraph 2002*, San Antonio, Texas.
[7] F. De la Torre, M. Black. "Robust parameterized component analysis: Theory and applications to 2D facial modeling". European Conf. On Computer Vision, ECCV 2002, Vol 4, pp. 653-669.
[8] M. Cohen, D.W. Massaro, "Modelling coarticulation in synthetic visual speech", Models and Techniques in Computer Animation, N.M. Thalmann and D.Thalmenn, Eds. Springer-Verlag, pp. 139-156
[9] M. Brand, "Voice Puppetry", *Proc. of the 26th annual conference on Computer graphics and interactive techniques*, p.21-28, July 1999.
[10] Dijkstra, E.W. A Note on Two Problems in Connexion with Graphs *In Numerische Mathematik vol. 1. 1959.*