APPEARS IN: IEEE MMSP'99, SEPTEMBER 13-15, 1999,

COPENHAGEN, DENMARK

ANALYSIS AND REPRODUCTION OF FACIAL EXPRESSIONS FOR COMMUNICATING CLONES

Stéphane Valente & Jean–Luc Dugelay
Institut Eurécom, Dpt of Multimedia Communications,
B.P. 193, 06904 Sophia–Antipolis Cedex, France

Tel.: +33 (0)4.93.00.26.{77,41}
Fax: +33 (0)4.93.00.26.27
E–mail: {valente,dugelay}@eurecom.fr

Abstract - **We present a novel view–based approach to quantify and reproduce facial expressions, by systematically exploiting the degrees of freedom allowed by a realistic face model, which embeds efficient mesh morphing and texture animations to synthesize facial expressions. For this purpose, we propose to use eigenfeatures, built from synthetic images, and design a linear estimator to interpret the responses of the eigenfeatures on a facial expression in terms of animation parameters.**

INTRODUCTION

Being able to analyze the facial expressions of a human face in a video sequence and reproduce them on a synthetic head model is of tremendous importance for many multimedia applications, like model–based coding, virtual actors, human–machine communication, interactive environments, video–telephony and virtual teleconferencing.

In the literature, three general analysis and animation techniques can be found to perform this task:

(i) **feature–based techniques and animation rules:** these methods are based on parametric face models, which are animated by a few parameters directly controlling the properties of facial features, like the mouth aperture and curvature, or the rotation of the eye–balls. The analysis technique consists in measuring some quantities on the user's face, for instance the size of the mouth area, using blobs, snakes or dot tracking. Some animation rules translate the measurements in terms of parametric animation parameters [1, 2, 3, 4];

(ii) **motion–based techniques and wireframe adaptation:** the motion information, computed on the user's face, is interpreted in terms of displacements of the face model wireframe, via a feedback loop. The face model can be either parametric or muscle–based [5, 6, 7];

(iii) **view–based techniques and key–frame interpolation:** the face animation is realized by interpolating the wireframe between several predefined configurations (*key–frames*), representing some extreme facial expressions. The difficulty of this approach is to relate the performer's facial expressions to the key–frames, and find the right interpolation coefficients. This is generally done by view–based techniques, which use appearance models of the distribution of pixel intensities around the facial features to characterize the facial expressions: in [8, 5], template–matching algorithms compute correlation scores with examples found in a database, and interpolation networks (a generalization of neural networks) produce the interpolation coefficients from the correlation scores, whereas [9] directly uses neural networks to estimate facial expressions from image pixels.

Although view–based techniques and key–frame interpolation are quite intuitive, they suffer from two difficulties. Firstly, the appearance models have to be carefully designed to take into account the coupling between the head pose (the $3D$ position and orientation of the user's face) and the facial expressions (for instance, if the performer nods his head downward, his mouth will be curved, and it could be interpreted as a smile), and the examples used to train the system must be closely related to the corresponding key–frames. Secondly, such a system is limited by the number of key–frames and training examples, as it will be difficult to analyze and reproduce a facial expression if it cannot be mimicked by a linear combination of key–frames. Needless to say, implementing these analysis and synthesis algorithms is a very empiric task.

In this paper, we assume we have a face model capable to produce facial images with a very high level of realism [10]. We propose a new view–based approach to relate the analysis and synthesis of facial expressions together, while solving the former limitations: a realistic face model, representing the human performer, is used to sample the visual space of facial expressions across various poses, via the face animation parameters. Then, a principal component analysis is performed over this training set, to extract a small set of images optimally spanning the training space. These images will allow us to characterize the facial expression of the user's face via a simple correlation mechanism. And finally, a linear estimator is designed to decouple the pose and expressions in the correlation scores, to relate the analysis to the face animation parameters. Such a system will not be limited by the amount of available key–frames, since all degrees of freedom permitted by the synthetic face are systematically exploited by the training strategy.

We also assume in the next discussion that we know the precise location of the facial features, and the global head pose: in [11], we presented an efficient analysis/synthesis feedback loop solving this issue for a video sequence.

SYNTHESIS OF FACIAL EXPRESSIONS

We build our face models from Cyberware^{TM} range data to obtain a realistic representation of the user [12]. They are made of a triangular wireframe, onto which a cylindrical texture is mapped, and preserve a level of complexity compatible with real–time manipulations. We implemented different animation techniques to generate flexible facial expressions [10]:

- mesh displacements, obtained by direct manipulations of the mesh vertices (see fig. 1(b));
- the gaze direction is controlled by drawing the pupils into the texture image (fig. 1(c));
- texture displacements, made by direct manipulations of the texture coordinates associated to the mesh vertices (fig. 1(d));
- texture blending, to alter the texture image at rendition time (fig. 1(e)).

(a) (b) (c) (d) (e)

Figure 1: Various animations: (a) neutral face model, resulting from the Cyberware^{TM} acquisition; (b) mesh displacements; (c) texture sliding; (d) texture displacements; (e) texture blending.

Following the guidelines defined by the MPEG–4 standard [13], the facial expressions of our face model are controlled by face animation parameters (FAP), which can gradually generate a given facial expression [14].

We can define an animation vector,

$$\boldsymbol{V} = (\boldsymbol{P}^T \boldsymbol{\mu}^T)^T = (t_x, t_y, t_z, r_x, r_y, r_z, \mu_1, \cdots, \mu_n)^T$$

which contains the head pose $\boldsymbol{P}$ and facial expression $\boldsymbol{\mu}$ parameters. The space of all possible facial expressions and head rotations is then sampled by

generating a set of $\boldsymbol{V}$ vectors, denoted $\{\boldsymbol{V}_i\}_{i=1,..,m}$, and the corresponding images are synthesized. Image patches for the facial features of interest (like the eyes, eyebrows, and the mouth) are extracted, to produce p datasets of training examples, denoted $\{\{\boldsymbol{D}_{i,j}\}_{i=1,..,m}\}_{j=1,..,p}$ where p is the number of considered facial features.

VISUAL MODELING OF FACIAL EXPRESSIONS

If we perform an eigendecomposition to reduce the number of images from m to q for each of the p datasets $\{\{\boldsymbol{D}_{i,j}\}_{i=1,..,m}\}_{j=1,..,p}$, we will obtain p sets of eigenfeatures $\{\boldsymbol{e}_{i,j}\}_{i=1,\cdots,q}$ which are optimal to decompose any image of $\{\boldsymbol{D}\}$ with the minimum square error between the image and its reconstruction. That is to say that for any facial expression for a given feature represented by image $\boldsymbol{I}$, there exists a vector $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_q)^T$ such as $\boldsymbol{I} \approx \overline{\boldsymbol{D}} + (e_1|\cdots|e_q)\boldsymbol{\lambda} = \overline{\boldsymbol{D}} + \boldsymbol{E}\boldsymbol{\lambda}$, where $\overline{\boldsymbol{D}}$ is the mean of the corresponding dataset.

The bases $\{\{\boldsymbol{e}_{i,j}\}_{i=1,\cdots,q}\}_{j=1,\cdots,p}$ are ideal to characterize a new facial expression in the sense that they exploit the visual redundancy of the training datasets to extract some compact and decorrelated parameters to represent facial expressions. As the eigenvectors are constructed for the model facial features, we can refer to them as *eigenfeatures*, capturing the pixel distribution in image patches due to both the face pose and facial expression.

ANALYSIS AND REPRODUCTION OF FACIAL EXPRESSIONS

Once the facial expressions are visually modeled by the previous eigendecomposition, a facial expression performed by the user, represented by image $\boldsymbol{I}$, is processed as follows (assuming the head pose has already been estimated): the facial features are correlated with the p bases of eigenfeatures, leading to the scores $\{\{\lambda_{i,j}\}_{i=1,\cdots,q}\}_{j=1,\cdots,p}$, which are concatenated along with the head rotation parameters into the vector

$$\boldsymbol{\lambda} = (r_x, r_y, r_z, \lambda_{1,1}, \lambda_{2,1}, \cdots, \lambda_{q-1,p}, \lambda_{q,p})^T$$

The problem now is to relate the vector $\boldsymbol{\lambda}$ to some vector $\boldsymbol{\mu}$ while decoupling the head pose from the facial expression. For this purpose, we construct the linear estimator $\boldsymbol{L}$, which satisfies the relation $\boldsymbol{\mu} = \boldsymbol{L}.\boldsymbol{\lambda}$ on the training database best in the least mean square sense. One can readily verify that this linear estimator is given by

$$\boldsymbol{L} = \boldsymbol{M}\boldsymbol{\Lambda}^T(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T)^{-1}$$

where $\boldsymbol{M} = (\boldsymbol{\mu}_1|\cdots|\boldsymbol{\mu}_d)$ and $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1|\cdots|\boldsymbol{\lambda}_d)$ are the matrices obtained by concatenating all $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ vectors from the training dataset.

EARLY RESULTS ON SYNTHETIC IMAGES

We first experimented this approach on synthetic images, to validate the analysis framework. A training dataset was created using the following simple sampling strategy: each FAP was altered one after another, taking the values $\{-1, -0.5, 0, 0.5, 1\}$, and for each obtained facial expression, we synthesized the face model under 9 different orientations, by setting the X and Y rotations to $\{-3, 0, 3\}$.

As it can be seen on figure 2, this approach works quite well for the animation of the eyes and eyebrows, and for expressions that are close to the training dataset in general, suggesting that the analysis strategy makes sense. However, it may have difficulties for some complicated expressions of the mouth, because many FAPs interact altogether in this area of the face, and the obtained facial expressions are too far from the training dataset, which is too simple for complex expressions.

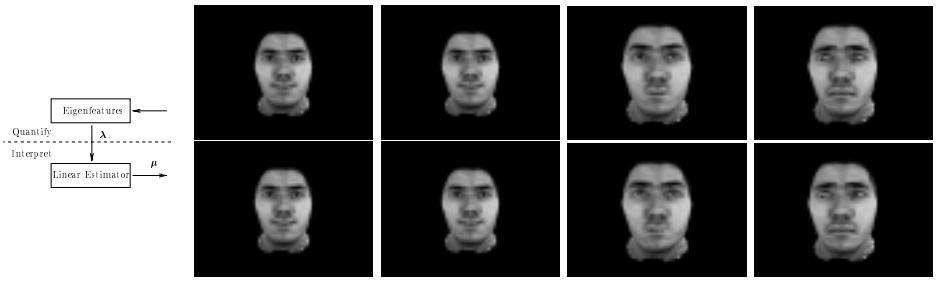

Figure 2: Some analyses of facial expressions: each image of the upper row (which does not belong to the training dataset) was quantified by some eigenfeatures, giving a $\boldsymbol{\lambda}$ vector. A linear estimator mapped $\boldsymbol{\lambda}$ to the animation parameters $\boldsymbol{\mu}$, which are rendered into the images of the lower row.

CONCLUDING REMARKS

We presented a novel view–based approach to quantify and reproduce facial expressions on a synthetic head model, by systematically exploiting the degrees of freedom allowed by a realistic face model. We proposed to use eigenfeatures, built from synthetic images, and designed a linear estimator to interpret the responses of the eigenfeatures on a facial expression in terms of animation parameters.

Our current perspectives include the refinement of the training strategy, and the extension of the algorithm on real facial images, to obtain a complete face cloning system [15].

ACKNOWLEDGMENTS

Eurécom's research is partially supported by its industrial members: Ascom, Cegetel, France Telecom, Hitachi, IBM France, Motorola, Swisscom, Texas Instruments, and Thomson CSF. The authors would like to thank Katia Fintzel (Espri Concept/Institut Eurécom) for being the model in figure 1.

References

[1] D. Terzopoulos. Modeling living systems for computer vision. In 14^{th} *International Joint Conference on Artificial Intelligence*, pages 1003–1013, Montreal, Quebec, August 1995.

[2] T. S. Huang, S. C. Reddy, and K. Aizawa. Human facial motion modeling, analysis and synthesis for video compression. *SPIE Vol. 1605 Visual Communications and Image Processing'91: Visual Communication*, pages 234–241, 1991.

[3] A. Saulnier, M.-L. Viaud, and D. Geldreich. Real–time facial analysis and synthesis chain. In *International Workshop on Automatic Face— and Gesture— Recognition*, pages 86–91, Zurich, Switzerland, 1995.

[4] J. Ohya, Y. Kitamura, F. Kishino, and N. Terashima. Virtual space teleconferencing: Real-time reproduction of tridimensional human images. *Journal of Visual Communication and Image Representation*, 6(1):1–25, March 1995.

[5] I. A. Essa, S. Basu, T. Darrell, and A. Pentland. Modeling, tracking, and interactive animation of faces and heads using input from video. In *Computer Animation'96 Conference*, Geneva, Switzerland, June 1996.

[6] P. Eisert and B. Girod. Facial expression analysis for model–based coding of video sequences. In *Picture Coding Symposium*, pages 33–38, Berlin, Germany, September 1997.

[7] H. Li, P. Roivainen, and R. Forchheimer. 3–D motion estimation in model–based facial image coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):545–555, June 1993.

[8] T. Darrell, I. Essa, and A. Pentland. Correlation and interpolation networks for real–time expression analysis/synthesis. Technical Report 284, MIT Media Lab.

[9] F. Hara and H. Kobayashi. An animated face robot — Its component technology development for interactive communication with humans. In *L'Interface des Mondes Réels & Virtuels*, Montpellier, France, Mai 1996.

[10] S. Valente and J.-L. Dugelay. Face tracking and realistic animations for telecommunicant clones. In *IEEE International Conference on Multimedia Computing and Systems*, Florence, Italy, June 7-11 1999.

[11] S. Valente and J.-L. Dugelay. 3D face modeling and encoding for virtual teleconferencing. In *Workshop on Very Low Bitrate Video (VLBV'98)*, Urbana–Champaign, Illinois, October 8-9 1998. In conjunction with ICIP'98.

[12] S. Valente, J.-L. Dugelay, and H. Delingette. Geometric and photometric head modeling for facial analysis technologies. Technical Report RR-98-041, Institut Eurécom, Sophia–Antipolis, France, May 1998.
URL http://www.eurecom.fr/~image/Publis98/RR-98-041.ps.gz

[13] Information technology — coding of audio–visual objects: Visual — ISO/IEC 14496–2. Comittee Draft ISO/IEC JTC1/SC29/WG11 N1902, International Organisation for Standardisation, Fribourg, Switzerland, October 1997.

[14] MPEG demo of the animation system.
URL http://www.eurecom.fr/~image/TRAIVI/animation.mpg

[15] J.-L. Dugelay, K. Fintzel, and S. Valente. Synthetic/natural hybrid video processings for virtual teleconferencing systems. In *Picture Coding Symposium*, Portland, Oregon, April 1999.