Shape and appearance models of talking faces for model-based tracking

M. Odisio G. Bailly

Institut de la Communication Parlée – CNRS, INPG, Université Stendhal
46, av. Félix Viallet - 38031 Grenoble Cedex 1 - France
{odisio,bailly}@icp.inpg.fr

Abstract

This paper presents a system that can recover and track the 3D speech movements of a speaker's face for each image of a monocular sequence. A speaker-specific face model is used for tracking: model parameters are extracted from each image by an analysis-by-synthesis loop. To handle both the individual specificities of the speaker's articulation and the complexity of the facial deformations during speech, speaker-specific models of the face 3D geometry and appearance are built from real data. The geometric model is linearly controlled by only six articulatory parameters. Appearance is seen either as a classical texture map or through local appearance of a relevant subset of 3D points. We compare several appearance models: they are either constant or depend linearly on the articulatory parameters. We evaluate these different appearance models with ground truth data.

1. Introduction

Intrinsic bimodality of speech is now well-known: in its production and its perception, speech consists of coherent audio and video signals. In presence of noise, audio-visual speech perception helps localization of the speaker (ventriloquist effect [6]) and enhances intelligibility, even for non-impaired people (cocktail party effect [9]). On the other hand, inconsistent audio and visual signals can perturb perception, increasing cognitive load [16] or even result in strong illusions: for instance, people confronted to a visual stimulus /aga/ synchronized with an audio stimulus /aba/ often perceive /ada/ (McGurk effect [15]). Also, face geometrical differences between two phonemes can be very subtle: /fu/ and /pu/ present a lips aperture difference of only a few mm^2. Finally, every speaker has his/her own articulation and speaking habits. A human interlocutor is very qualified for noticing and interpreting these specificities, as well as inconsistencies, if any.

For most applications for communication that involve virtual talking faces, high fidelity is required when resynthesizing facial movements, and thus for their prior extraction. Tracking speech movements in a video is a challenging task because expected reconstruction results must be very accurate.

In this paper, we have in mind virtual teleconferencing applications: in a common virtual space, every participant is represented by a 3D delegate reproducing the gestures of his/her owner. We face the problem of robustly recovering the 3D speech movements of a given speaker from monocular images.

To achieve such tasks, making use of a 3D face model is a very popular approach. Generic 3D models can be classified as parametric [17][21] or physics-based [23][3]. These *a priori* models must be customized to the anatomy of the speaker before tracking his/her facial movements. Even then, it is not guaranteed that they could be fairly adapted to every facial configurations. Data driven models can cope with this problem [10][24][11]. We will describe below a methodology that lets control parameters of a speaker-specific model emerge from a statistical analysis of fine-grained 3D data.

With only a 3D model, low-level image processing techniques are usually employed to extract features such as interest points, gradient, or edges [4][22]. Then, these 2D measurements must be inverted to determine the control parameters of the 3D model: this operation may be an *ill-posed* problem as the solution may not exist, or may not be unique. Because optical flow generates a large number of correspondences, the inversion is more likely to lead to a solution [13]. It can be combined with edge-adjustment [5], or with analysis-by-synthesis technique [7].

Face images depend on head motion, illumination conditions and facial movements. In presence of large image changes, tracking can take great advantage of an appearance-variations model, which is moreover included in an analysis-by-synthesis loop. This has been successfully applied for recovering head pose [12], expressions [18], or identity [20].

As we concentrate on speech movements, we assume small head motion and small illumination variations. Following our 3D modelling methodology, we use statistical analysis to build the appearance models. They are linearly controlled by the same articulatory parameters that drive the geometric model. In addition to classical texture mapping of the whole face, model of local appearance of a subset of relevant 3D points is also implemented.

Our 3D geometric model is presented in next section. The appearance models are described in section 3. Then, the tracking stage is detailed and finally results of several experiments are discussed.

2. Geometric model

The geometric model of the speaker's facial movements is 3D, linear and driven by a vector α of six articulatory parameters. A translation $\mathbf{t}$ and a rotation $\mathbf{R}$ define the global head motion which frames the speech movements. Finally, the 3D face model is entirely controlled by the set of parameters p:

$$p = [\ \alpha^T \quad t_x \quad t_y \quad t_z \quad r_x \quad r_y \quad r_z\ ]^T \qquad (1)$$

We follow the methodology of [1]; constructing an articulatory model specific to the speaker so as to capture its audiovisual speech activity comprises two stages: collecting accurate 3D data of the speaker and analysing them through a statistical iterative scheme.

2.1. Data acquisition

As shown in figure 1, glued coloured beads mark a few hundreds fleshpoints all over the speaker's face. Using calibrated cameras and mirrors, we record a few dozen visemes — typical articulatory configurations of the language — on which each bead is semi-automatically labeled on each of the four views. The 3D coordinates of each fleshpoint and of 30 points characterizing lips shape [1] are then reconstructed and expressed in a referential linked to the bite plane.

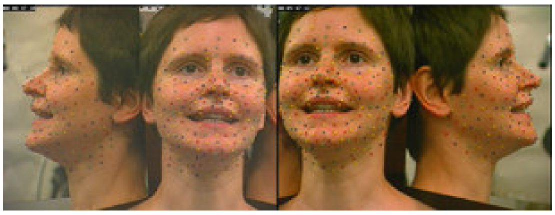

Figure 1: Data acquisition setup. Speaker uttering /iki/

2.2. Data modelling

The 3D linear model emerges from iterative statistical analysis of these 3D data [1]. At each successive step of the model construction, a Principal Component Analysis is performed on an *a priori*-selected subset (coordinates of jaw, lips, etc.) of the residual data; the main directions of this PCA are retained as *linear predictors* for the whole residual data. Facial speech movements are hence simply modelled:

$$X = [\, x_1\, y_1\, z_1\ \ldots\ x_n\, y_n\, z_n\,]^T = X_0 + \mathbf{M_X} \cdot \alpha \qquad (2)$$

With only six (non-linearly correlated) parameters α, the model $\mathbf{M_X}$ explains more than 95% of the data variance. These six elementary degrees of freedom include jaw opening, lips rounding (see figure 2), or lips opening but also movements such as lip raising, or jaw advance. By linear addition, they allow to reproduce trustfully the geometry of the visemes: for example, labiodentals (*e.g.* /v/, /f/) require both retracting the jaw and pulling up both lips to ensure contact between the lower lip and the upper teeth; palatal fricatives (*e.g.* /ʒ/, /ʃ/) require both lip rounding and large aperture.

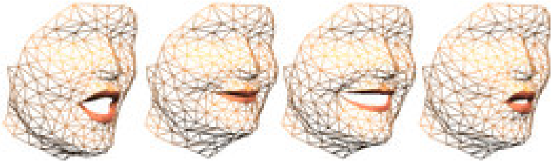

Figure 2: Two elementary movements of the model. Open/close jaw and spread/round lips

3. Appearance models

Thanks to the accuracy of the geometric model, it is likely that, for fine 3D tracking purpose, the appearance model needs only to bring *few information* so that the tracking algorithm produces satisfying results. As we assume small head motion and small illuminating changes, applying here widely used techniques based on appearance eigenspaces [24][20][26] would just complexify the task: it would add to the face model appearance parameters that would have to be mapped to the shape parameters.

The two different appearance models detailed below depend linearly on the articulatory parameters; linked to shape changes, appearance changes can be reinterpreted into facial movements.

3.1. Texture mapping (tex)

Assuming an image, a texture I can be defined by a set of geometric parameters p. Warping the geometry to a given posture (corresponding to p_0) allows to warp I to a normalized-shape image referential. Its dimensions are constant and depend only on p_0. In this referential, a variable texture is modelled as:

$$I_{tex} = [\,R_1\,G_1\,B_1\;\ldots\;R_m\,G_m\,B_m\,]^T = I_0 + \mathbf{M_I}\cdot\alpha \quad (3)$$

For synthesis of the face, I_{tex} is computed and served as texture for the morphed 3D posture. Projection onto the image plane finally leads to a set $\mathcal{S}$ of *rendered* pixels.

As an example, such a model was learned on visemes used for the geometric modelling. Due to the variations of head motion, slight light variations and the warping noise, it explains only 50% of the data variance. However, as illustrated in figure 3 (and in figure 4 for tracking) it renders properly the major appearance changes, such as the different aspects of the nasogenian wrinkle for spread or rounded lips.

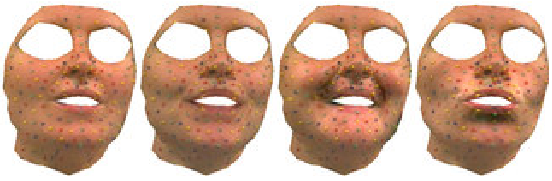

Figure 3: The reference shape textured with different synthesized normalized-shape textures

3.2. Model of local appearance (la)

A marker-free face contains large parts where the texture is very poor and not subject to major variations. It seems then interesting to model appearance only for the more informative regions. We do so by modelling the local appearance of selected 3D points.

We describe the local appearance with a vector d containing responses to Gaussian derivative filters [14]. For a 3D point, the convolutions are computed at its projection (u, v) on the image I. To ease the dissociation between luminance and chrominance, image values $I(u, v)$ are expressed in the colour space (Y, C_b, C_r):

$$\begin{pmatrix} Y \\ C_b \\ C_r \end{pmatrix} = \begin{pmatrix} 0.2220 & 0.7067 & 0.0713 & 0 \\ -0.1195 & -0.3805 & 0.5000 & 127.5 \\ 0.5000 & -0.4542 & -0.0458 & 127.5 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \\ 1 \end{pmatrix}$$

To approximate the local image variation, up to the first Gaussian derivatives are represented in d; the zeroth order derivative of the Y channel is discarded so that d is not sensitive to luminance offset changes; this leaves eight components for d:

$$d = \begin{pmatrix} Y \\ C_b \\ C_r \end{pmatrix} \star \left(G_0(\sigma_0) \quad G_1^x(\sigma_0) \quad G_1^y(\sigma_0)\right) \setminus \left\{Y_{G_O(\sigma_0)}\right\}$$

where the convolutions are computed at a scale σ_0, which is kept constant in this paper.

Again, for each point P^i of the 3D model, d^i is linearly modelled by the articulatory parameters:

$$d^i = d_0^i + \mathbf{M_D^i}\cdot\alpha \qquad (4)$$

For each articulatory parameter, a few points are automatically selected according to the variance of the data reconstructed by the corresponding column of $\mathbf{M_D^i}$. Duplicate are removed when merging into the final set.

Finally, appearance of the face is defined as a set of N 3D points local appearance:

$$\mathbf{D} = \begin{bmatrix} d^1 & d^2 & \ldots & d^N \end{bmatrix} \qquad (5)$$

Texture-mapping appearance model can also be formalized in (5): then, N is the cardinal of $\mathcal{S}$, and each descriptor d^i is a vector containing the luminance-normalized RGB values (*ie.* divided by $L = R + G + B$) of the i^{th} pixel of $\mathcal{S}$.

4. Tracking algorithm

The face model described in sections 2 and 3 synthesizes a set of appearance descriptors $\mathbf{D^s}$ that corresponds to a vector of control parameters p. Our fitting algorithm performs the inverse task: it aims to recover the parameters $\hat{p}$ which synthesizes the descriptors $\hat{\mathbf{D}}^{\mathbf{s}}$ that best match the descriptors $\mathbf{D^a}$ of the analysed image I^a. The dissimilarity between parameters p and image I^a is measured as the distance between the descriptors $\mathbf{D^s}$ and $\mathbf{D^a}$:

$$\varepsilon(p) = \frac{1}{NN_d}\sum_{i=1}^{N}\sum_{j=1}^{N_d}\left(D_{i,j}^a(p) - D_{i,j}^s(p)\right)^2 + A(\alpha)$$

where N_d-coordinates $D_{i,j}^a(p)$ and $D_{i,j}^s(p)$ are computed according to section 3, and $A(\alpha)$ is an exponentially-fashioned function that penalizes improbable articulatory parameters.

The purpose of our analysis-by-synthesis optimization scheme is to deliver: $\hat{p} = \arg\min_p \varepsilon(p)$

We have compared several classical optimization methods, including Levenberg-Marquardt and local variations. The best convergence results were obtained by the Nelder-Mead downhill simplex algorithm [19].

For sequence tracking, the best fitting parameters for a given image are used as the initial simplex centroid for the following one.

5. Experimental results

The goal of the experiments detailed below is both to evaluate our system and to compare the different appearance models on *real* ground truth data. These ground truth data were obtained by computing a direct inversion of the geometric model from semi-automatically labelled positions of glued beads. Cameras were calibrated in all experiments.

5.1. Validation experiments

For the following experiments, the tracking system was evaluated in the geometric modelling conditions. Out of the four views of the setup (see figure 1), only one front view was used.

The visemes used in constructing the geometrical model have been randomly separated in two classes: 44 visemes have been dedicated to constructing the appearance models, and 11 visemes have been kept apart for the tests.

We have built four appearance models: texture mapping models tex_lin and tex_cst are controlled according to (3), tex_cst doesn't vary with α ($\mathbf{M_I} = \mathbf{0}$); similarly, the models of local appearance la_lin and la_cst are controlled according to (4), la_cst doesn't vary with α ($\mathbf{M_D^i} = \mathbf{0}$). They use $N = 63$ 3D points.

The first tracking experiment was performed on the tests visemes. Head motion was precisely estimated during the construction of the geometric model, and only articulatory parameters were tracked, using the neutral posture as the initial conditions. Figure 4 shows the residual 3D error computed including all the 3D points of the geometric model. With appearance models tex_lin, tex_cst and la_lin the system quite successfully recovers the geometry of the visemes. A large part of the visemes have residual 3D error below the uncertainty of the geometric modelling. Results with tex_lin and la_lin are better than those with tex_cst and la_cst, showing the benefit of the articulatory-dependent modelling. Some failures, such as for /asa/, have occurred; however, the tracking was initialised far from the solution: when tracking a sequence, only few variations have to be estimated between two successive frames.

Another evaluation was performed by tracking on a complete video sequence both head motion and articulatory posture. As we can see on the figure 5, best results are obtained with tex_lin; tex_cst and la_lin behave equally well; la_cst is clearly worse. The very simple geometric measurements lip width and lip aperture allow to see that the estimated articulatory movements reproduce the anticipatory gestures and reach *phonetic* targets such as lips closure for the bilabial stops /p/ and /m/. On average, the error function is called around 300 times per frame upon convergence, mainly because we have kept the very conservative ending criteria used when tracking the visemes.

Here, the beads glued on the whole speaker's face enhance texture details; this bias yields great advantage to the texture-mapping models.

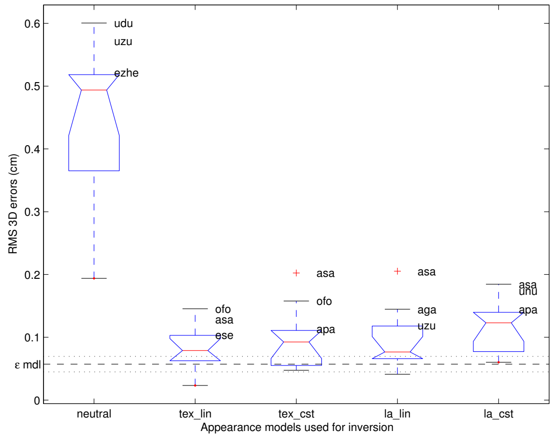

Figure 4: Boxplots [1] of the tracking residual RMS 3D error on test visemes for different appearance models. For each group, the worst three visemes are labelled. For comparison, 'neutral' shows the initial error.

5.2. Teleconferencing experiment

We now consider conditions that could be those of a real teleconference: the speaker is filmed by a single head-mounted calibrated micro-camera and his/her face is marked with only a few beads sufficient to compute reliably a direct inversion of the articulatory model. The main region of interest — the lips — is left unmarked.

Data from a whole sequence was used for constructing the appearance models. We have only built the following models (denoted as above): tex_cst (first image of the sequence) and la_lin (using a whole 75 images sequence). They can both be computed very fast, either by graphical hardware for tex_cst or by software for la_lin.

For the model of local appearance la_lin, 3D points in the neighborhood of a bead have been removed of the automatic selection process. The result shown in figure 6 makes sense: the $N = 51$ retained points are mainly distributed on lips, throat and on the jaw line, including also a point located at the beginning of the nasogenian wrinkle.

These two appearance models have been tested by tracking the same sequence as used for the models constructions (*cf.* figure 7). In this experiment, results with la_lin are much better than those with tex_cst. Average tracking residual RMS 3D error is 0.25 cm for tex_cst and 0.13 cm for la_lin. However, presence of peaks during the sequence indicate that even with the model of local appearance the tracking system could fail on a few frames (see figure 8).

[1] A boxplot provides a visual description of several statistical aspects of a data sample.

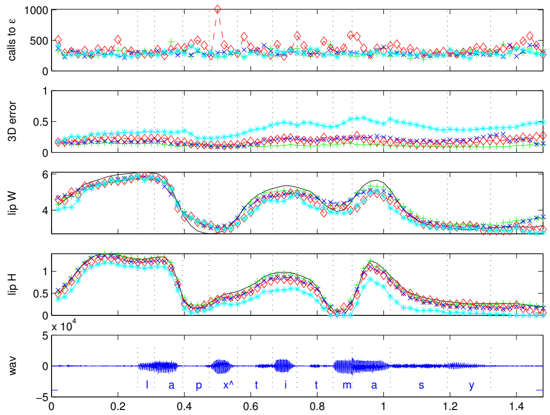

Figure 5: Tracking of the sequence “massue” with different appearance models. tex_lin(’+’), tex_cst(’x’), la_lin(’◇’), la_cst (’*’). Ground truth data is the solid line. From top to bottom: number of evaluations of ε, RMS 3D error (cm), lip width (cm), lip aperture (cm) and labelled audio.

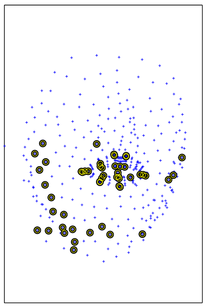

Figure 6: Frontal view of the 3D points showing in big circles the points automatically selected for the model of local appearance on the corpus “annie”.

6. Conclusions

We have presented an original system to estimate the 3D speech movements of a speaker’s face from a video sequence. A speaker-specific face model is used for tracking: model parameters are extracted from each image by an analysis-by-synthesis loop. To capture the individual specificities of the speaker’s articulation, an accurate 3D model of the face geometry and an appearance model are built from real data. The geometric model is linearly controlled by only six articulatory parameters. We have compared several appearance models, where appearance is seen either as a classical texture map or through local appearance of an automatically selected subset of 3D points. Evaluation with ground truth data has shown satisfying results for the texture mapping models and for the model of local appearance linearly controlled by

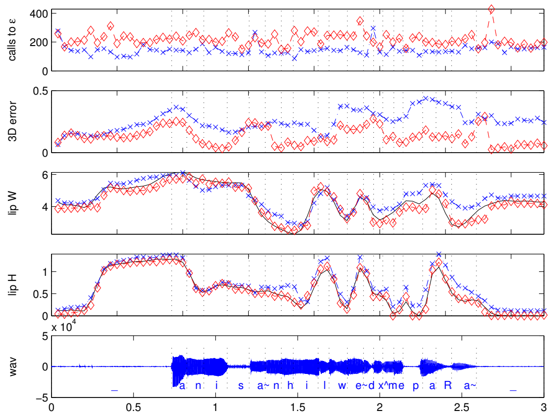

Figure 7: Tracking of the sequence “annie” with different appearance models. tex_cst (’x’), la_lin (’◇’). Ground truth data is the solid line. From top to bottom: number of evaluations of ε, RMS 3D error (cm), lip width (cm), lip aperture (cm) and labelled audio.

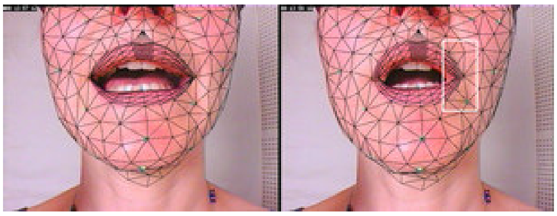

Figure 8: Examples of correct (left) and incorrect (right) la_lin model adjustment on two images of the sequence “annie”. The box highlights mismatch in lip left corner.

the articulatory parameters. This latter point allows to believe that our tracking system could reach real-time in a foreseeable future. Moreover, it can be completed with a module mapping facial speech movements to standardized MPEG-4 FAP [8] to form an efficient model-based encoder.

Further work will include extending the abilities of our tracking system and subjective evaluation.

To compute the local appearance, the Gaussian filters box is *rigidly* 2D. Deforming it by taking into account the corresponding 3D surface as in [25] is a step toward the construction of a new model of local appearance that would span several illuminating conditions. Also, each point should be considered at its intrinsic scale.

When tracking a sequence, a module for temporal prediction of the control parameters that includes audio information is under development.

Eventually, subjective evaluation [2] of our tracking results by intelligibility tests will provide more clues for understanding how our virtual talking heads are perceived.

Acknowledgments

We thank F. Elisei, P. Badin and B. Holm for their valuable input, and H. Lœvenbruck as the subject of this study.

References

[1] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images," *Journal of Phonetics*, vol. 30, no. 3, pp. 533–553, 2002.

[2] G. Bailly, G. Gibert, and M. Odisio, "Evaluation of movement generation systems using the point-light technique," in *IEEE Workshop on Speech Synthesis*, Santa Monica, USA, Sept. 2002.

[3] S. Basu, N. Oliver, and A. Pentland, "3D lip shapes from video: A combined physical-statistical model," *Speech Communication*, vol. 26, pp. 131–148, 1998.

[4] L. Bretzner and T. Lindeberg, "Qualitative multi-scale feature hierarchies for object tracking," in *Proceedings of the International Conference on Scale-Space Theories in Computer Vision*, ser. Lecture Notes in Computer Science, vol. 1682. Corfu, Greece: Springer-Verlag, Sept. 1999, pp. 117–128.

[5] D. DeCarlo and D. Metaxas, "Optical flow constraints on deformable models with applications to face tracking," *International Journal of Computer Vision*, vol. 38, no. 2, pp. 99–127, July 2000.

[6] J. Driver, "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading," *Nature*, vol. 381, no. 6577, pp. 66–68, May 1996.

[7] P. Eisert and B. Girod, "Analyzing facial expressions for virtual conferencing," *IEEE Computer Graphics & Applications*, vol. 18, no. 5, pp. 70–78, Sept. 1998.

[8] F. Elisei, M. Odisio, G. Bailly, and P. Badin, "Creating and controlling video-realistic talking heads," in *Auditory-visual Speech Processing Workshop*, Aalborg, Denmark, Sept. 2001, pp. 90–97.

[9] N. P. Erber, "Interaction of audition and vision in the recognition of oral speech stimuli," *Journal of Speech and Hearing Research*, vol. 12, pp. 423–425, 1969.

[10] B. Guenter, C. Grimm, D. Wood, H. Malvar, and F. Pighin, "Making faces," in *Proceedings of SIGGRAPH*, ser. Computer Graphics Proceedings, Annual Conference Series. ACM SIGGRAPH / Addison Wesley, July 1998, pp. 55–66.

[11] P. Hong, Z. Wen, and T. S. Huang, "Real-time speech-driven face animation," in *MPEG-4 Facial Animation. The Standard, Implementation and Applications.*, I. S. Pandzic and R. Forchheimer, Eds. Wiley, 2002, ch. 7, pp. 115–124.

[12] M. La Cascia, S. Sclaroff, and V. Athitsos, "Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3D models," *IEEE Transations on Pattern Analysis and Machine Intelligence*, vol. 22, no. 4, pp. 322–336, Apr. 2000.

[13] H. Li, P. Roivainen, and R. Forchheimer, "3-D motion estimation in model-based facial image coding," *IEEE Transations on Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 545–555, June 1993.

[14] T. Lindeberg, "Feature detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 77–116, 1998.

[15] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, Dec. 1976.

[16] I. S. Pandzic, J. Ostermann, and D. Millen, "User evaluation: synthetic talking faces for interactive services," *The Visual Computer*, vol. 15, pp. 330–340, 1999.

[17] F. I. Parke, "Parameterized models for facial animation," *IEEE Computer Graphics & Applications*, vol. 2, pp. 61–68, Nov. 1982.

[18] F. Pighin, R. Szeliski, and D. H. Salesin, "Modeling and animating realistic faces from images," *International Journal of Computer Vision*, vol. 50, no. 2, pp. 143–169, Nov. 2002.

[19] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed. Cambridge University Press, 1992.

[20] S. Romdhani, V. Blanz, and T. Vetter, "Face identification by fitting a 3D morphable model using linear shape and texture error functions," in *Proceedings of the European Conference on Computer Vision*, ser. Lecture Notes in Computer Science, vol. 2353. Copenhagen, Denmark: Springer-Verlag, May 2002, pp. 3–19.

[21] M. Rydfalk, "CANDIDE, a parameterized face," Dept. of Electrical Engineering, Linköping University, Tech. Rep. LiTH-ISY-I-866, 1987.

[22] J. Ström, T. Jebara, S. Basu, and A. Pentland, "Real time tracking and modeling of faces: an EKF-based analysis by synthesis approach," in *Proceedings of International Conference on Computer Vision*, Corfu, Greece, Sept. 1999.

[23] D. Terzopoulos and K. Waters, "Analysis and synthesis of facial image sequences using physical and anatomical models," *IEEE Transations on Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 569–579, June 1993.

[24] B. J. Theobald, J. A. Bangham, I. Matthews, and G. C. Cawley, "Visual speech synthesis using statistical models of shape and appearance," in *Auditory-visual Speech Processing Workshop*, 2001, pp. 78–83.

[25] C. S. Wiles, A. Maki, and N. Matsuda, "Hyperpatches for 3D model acquisition and tracking," *IEEE Transations on Pattern Analysis and Machine Intelligence*, vol. 23, no. 12, pp. 1391–1403, Dec. 2001.

[26] S. Yan, C. Liu, S. Li, H. Zhang, H. Shum, and Q. Cheng, "Texture-constrained active shape models," in *Proceedings of The First International Workshop on Generative-Model-Based Vision*, Copenhagen, Denmark, May 2002.