Published in the Proceedings of the *6th IEEE International Conference on Automatic Face and Gesture Recognition (FG'04)*, Seoul, Korea, May, 2004, 3–8.

Model-Based and Image-Based Methods for Facial Image Synthesis, Analysis and Recognition

Demetri Terzopoulos[1,2], **Yuencheng Lee**[2,1] and **M. Alex O. Vasilescu**[2,1]

[1]Courant Institute of Mathematical Sciences, New York University, New York, NY 10003, USA
[2]Department of Computer Science, University of Toronto, Toronto ON M5S 3G4, Canada

Abstract

We review several model-based and image-based methods that we have developed for analyzing, synthesizing, and recognizing facial images. Our model-based methods include a sophisticated, functional model of the human face/head, which incorporates a biomechanical tissue model with embedded muscle actuators, and techniques for applying it to computer animation and expression estimation in video. Our image-based methods include TensorFaces, a nonlinear (multilinear) representation for facial image ensembles that disentangles pose, illumination, and expression effects to improve facial recognition.

1. Introduction

After more than three decades of intense research, the human face continues to pose challenging research problems in computer vision, computer graphics, and related fields. Computer vision researchers have been mainly interested in tackling facial image analysis and recognition problems (see, e.g., [2]). Conversely, computer graphics researchers have focused on problems of facial image synthesis (see, e.g., [7]). In both domains, interest extends beyond static images to video. When it comes to faces, vision researchers have traditionally relied on image-based methods while their graphics colleagues have traditionally relied on model-based methods, but this is no longer strictly the case. Strongly model-based and image-based methods may be regarded as being situated at opposite ends of a spectrum, with myriad hybrid approaches in between, most of them yet to be conceived.

Within our research group, we have had a longstanding interest both in computer vision and in computer graphics, including subinterests in facial image/video analysis, recognition, and synthesis. This article overviews several of our contributions on these topics. The remainder of the article is divided into two main sections and a conclusion. Section 2 reviews our model-based methods, which exploit a sophisticated, function model of the human face, and their application to the synthesis of animated facial expressions and in the analysis of expressive facial video. We also discuss the individualization of our face model to fit image sensor data acquired from human subjects. We then turn in Section 3 to nonlinear image-based approaches for face recognition, reviewing our recently proposed appearance-based recognition method known as TensorFaces, which relies on multilinear algebra, the algebra of higher-order tensors.

2. Model-Based Methods

2.1. A Functional Facial Model

We have developed a sophisticated, functional model of the human face and head that is efficient enough to run at interactive rates on high-end PCs. Conceptually, the model decomposes hierarchically into several levels of abstraction, which represent essential aspects related to the psychology of human behavior and facial expression, the anatomy of facial muscle structures, the histology and biomechanics of facial tissues, facial geometry and skeletal kinematics, and graphical visualization:

1. *Behavior.* At the highest level of abstraction, the synthetic face model has a repertoire of autonomous behaviors, including natural head/eye behaviors, as well as intentional and reactive expressive behaviors.

2. *Expression.* At the next level, the face model executes individual expression commands. It can synthesize any of the six primary expressions within a specific duration and degree of emphasis. A muscle control process based on Ekman and Friesen's FACS [2] translates expression instructions into the appropriately coordinated activation of actuator groups in the soft-tissue model. This coordination offers a semantically rich set of control parameters which reflect the natural constraints of real faces.

3. *Muscle Actuation.* As in real faces, muscles comprise the basic actuation mechanism of the face model. Each muscle submodel consists of a bundle of muscle fibers. Currently there are some three dozen muscles of facial expression in the synthetic face.

(a) (b) (c)

Figure 1: Image-based facial modeling. (a) Cylindrical range and texture images of the head of a real person captured using a Cyberware 3D Color Digitizer. From the pair of images at the top, our algorithms "clone" a functional model of the subject, incorporating a textured, biomechanically-simulated deformable facial skin with embedded muscles of facial expression. The synthetic face at the bottom is rendered in neutral (center) and expressive poses dynamically generated through coordinated muscle contractions. (b) Fitting the generic mesh to both RGB texture and edge-enhanced range images. (c) Scenes from the computer-animated short *Bureaucrat Too*, which features an animated face "cloned" from a male subject.

4. *Biomechanics.* When muscle fibers contract, they displace their points of attachment in the facial tissue or the articulated jaw. The face model incorporates a physical approximation to human facial tissue, a non-homogeneous and nonisotropic layered structure consisting of the epidermis, dermis, subcutaneous fatty tissue, fascia, and muscle layers. The tissue model [6] is a lattice of point masses connected by nonlinear viscoelastic springs, arranged as layered prismatic elements that are constrained to slide over an impenetrable skull substructure. Large-scale synthetic tissue deformations are simulated numerically by continuously computing the response of the assembly of volume-preserving elements to the stresses induced by activated muscle fibers.

5. *Geometry/Kinematics.* The geometric representation of the facial model is a non-uniform mesh of polyhedral elements whose sizes depend on the curvature of the neutral face. Muscle-induced synthetic tissue deformations distort the neutral geometry into an expressive geometry. The epidermal display model is a smoothly-curved subdivision surface that deforms in accordance with the simulated tissue elements. In addition, the complete head model includes functional subsidiary models of skull with articulated jaw, teeth, tongue/palate, eyes, eyelids, and neck.

6. *Rendering.* After each simulation time step, standard visualization algorithms implemented in the PC graphics pipeline render the deforming facial geometry in accordance with viewpoint, light source, and skin reflectance (texture) information to produce the lowest level representation in the modeling hierarchy, a continuous stream of facial images.

The hierarchical structure of the model appropriately encapsulates the complexities of the underlying representations, relegating the details of their simulation to automatic procedures.

2.2. Image-Based Reconstruction of the Face Model

We have developed a highly automated image-based approach to constructing anatomically accurate, functional models of human heads that can be made to conform closely to specific individuals [6]. Fig. 1(a) shows example input images and the resulting functional model, which is suitable for animation. The image acquisition phase begins by scanning a human subject with a laser sensor, which circles around the subject's head to acquire detailed range and reflectance images. The figure shows a head-to-shoulder, 360° cylindrical scan of a woman, "Heidi", acquired using a Cyberware Color 3D Digitizer, producing a range image and a registered RGB photometric image, both 512×256 pixel arrays in cylindrical coordinates.

In the image analysis phase, an automatic conformation algorithm adapts to the acquired images a deformable model which takes the form of an elastic triangulated face mesh of predetermined topological structure. The generic mesh, which is reusable with different individuals, reduces the range data to an efficient, polygonal approximation of the facial geometry and supports a high-resolution texture mapping of the skin reflectivity. Fig. 1(b) shows the elastic mesh after it has conformed to the woman's facial area in both the range and RGB images using a feature-based matching algorithm that encodes structural knowledge about the face, specifically the relative arrangement of nose, eyes, ears, mouth, and chin. The 2D positions of the nodes of the conformed mesh serve as texture map coordinates in the RGB image, as well as range map sampling

locations from which 3D Euclidean space coordinates are computed for the polygon vertices. The visual quality of the face model is comparable to a 3D display of the original high resolution data, despite the significantly coarser mesh geometry.

After reducing the scanned data to the 3D epidermal mesh, the final phase of the reconstruction process assembles the physics-based, functional face model. The conformed polygonal mesh forms the epidermal layer of a biomechanical model of facial tissue. An automatic algorithm constructs the multilayer synthetic skin and estimates an underlying skull substructure with a jointed jaw. Finally, the algorithm inserts the synthetic muscles of facial expression into the deepest layer of the facial tissue. The resulting face model can be animated, as is illustrated for a male subject in Fig. 1(c).

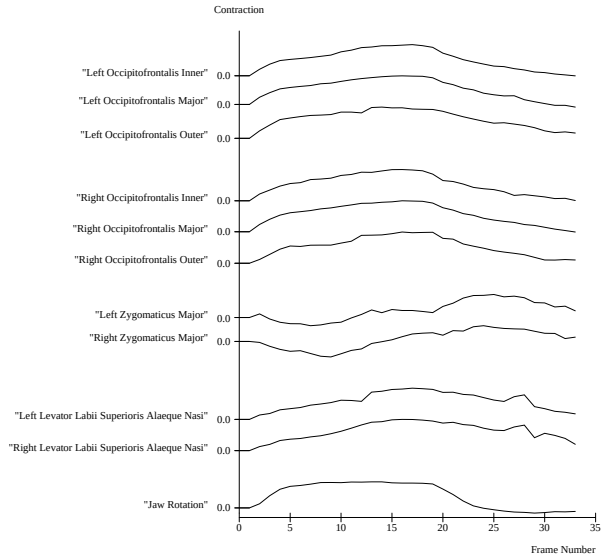
2.3. Model-Based Facial Image Analysis/Synthesis

Facial image analysis/synthesis is useful in several applications. Among them is low bandwidth teleconferencing which may involve the real-time extraction of facial control parameters from live video at the transmission site and the reconstruction of a dynamic facsimile of the subject's face at a remote receiver. Teleconferencing and other applications require facial models that are computationally efficient and also realistic enough to synthesize the various nuances of facial structure and motion. Over a decade ago, we argued that the anatomy and physics of the human face, especially the arrangement and actions of the primary facial muscles, provide strong constraints and a principled basis for facial image analysis and synthesis [8].

Part of the difficulty of facial image analysis is that the face is highly deformable, particularly around the forehead, eyes, and mouth, and these deformations convey a great deal of meaningful information. Techniques for tracking the deformation of facial features include "snakes" [5]. Motivated by the anatomically consistent musculature in our model, we have considered the estimation of dynamic facial muscle contractions from video sequences of expressive faces (e.g., 2(a)). We have developed an analysis technique that uses snakes to track the nonrigid motions of facial features in video (2(b)). Features of interest include the eyebrows, nasal furrows, mouth, and jaw in the image plane. We are able to estimate dynamic facial muscle contractions directly from the snake state variables. Fig. 2(d) shows a plot of the estimated muscle contractions versus the frame number. They are input at the muscle actuation layer of the functional model. These estimates make appropriate control parameters for resynthesizing facial expressions through a generic face model at real-time rates (2(c)). Three rendered images are shown in Fig. 2(c).

Ishikawa et al. [4, 3] present a variant on this approach, using a neural network transducer to map between estimated muscle actions and the synthetic face model.

(a) (b) (c)



(d)

Figure 2: Dynamic facial image analysis and expression resynthesis. Sample video frames with superimposed deformable contours tracking facial features; (a) intensity images with black snakes, (b) image potentials with white snakes. (c) Facial model resynthesizes surprise expression from estimated muscle contractions. (d) Estimated facial muscle contractions plotted as time series.

3. Image-Based Methods

Appearance-based face recognition has been an active area of biometrics research in recent years [1]. Given a database of suitably labeled training images of numerous individuals, this supervised pattern-recognition technique aspires either to recognize the faces of these individuals in previously unseen test images or to identify the test images as new faces. The conventional approach addresses the problem of facial representation for recognition by taking advantage of the functionality and simplicity of linear algebra, the algebra of matrices. In particular, principal components analysis (PCA) has been a popular method for appearance-based facial image recognition. This linear method (a.k.a. "eigenfaces") and its variants adequately address face recognition under tightly constrained conditions—e.g., frontal mugshots, fixed lightsources, fixed expression—where person identity is the only factor that is allowed to vary. Various attempts have been made to deal with the shortcomings of PCA-based facial image representations in less constrained situations.

In our appearance-based recognition work, we confront the fact that natural images result from the interaction of *multiple* factors related to scene structure, illumination, and imaging. For facial images, these factors include different facial geometries, expressions, head poses, and lighting conditions. We have advocated the use of multilinear algebra, the algebra of higher-order tensors, for computing a parsimonious representation of facial image ensembles which separates these factors [10]. Our representation, called *TensorFaces*, yields significantly improved facial recognition rates relative to standard eigenfaces [9].

3.1. TensorFaces

Within the TensorFaces framework, the image ensemble is represented as a higher-dimensional tensor. This image data tensor $\mathcal{D}$ must be decomposed in order to separate and parsimoniously represent the constituent factors related to scene structure, illumination, and viewpoint. To this end, we prescribe the N*-mode SVD* algorithm, a multilinear extension of the conventional matrix singular value decomposition (SVD). Our earlier papers overview the mathematics of our multilinear analysis approach and presents the N-mode SVD algorithm [10].

In short, an order $N > 2$ tensor or N-way array $\mathcal{D}$ is an N-dimensional matrix comprising N spaces. N-mode SVD is a "generalization" of conventional matrix (i.e., 2-mode) SVD. It orthogonalizes these N spaces and decomposes the tensor as the *mode-n product*, denoted $\times_n$, of N-orthogonal spaces, as follows:

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \ldots \times_n \mathbf{U}_n \ldots \times_N \mathbf{U}_N. \quad (1)$$

Tensor $\mathcal{Z}$, known as the *core tensor*, is analogous to the diagonal singular value matrix in conventional matrix SVD,

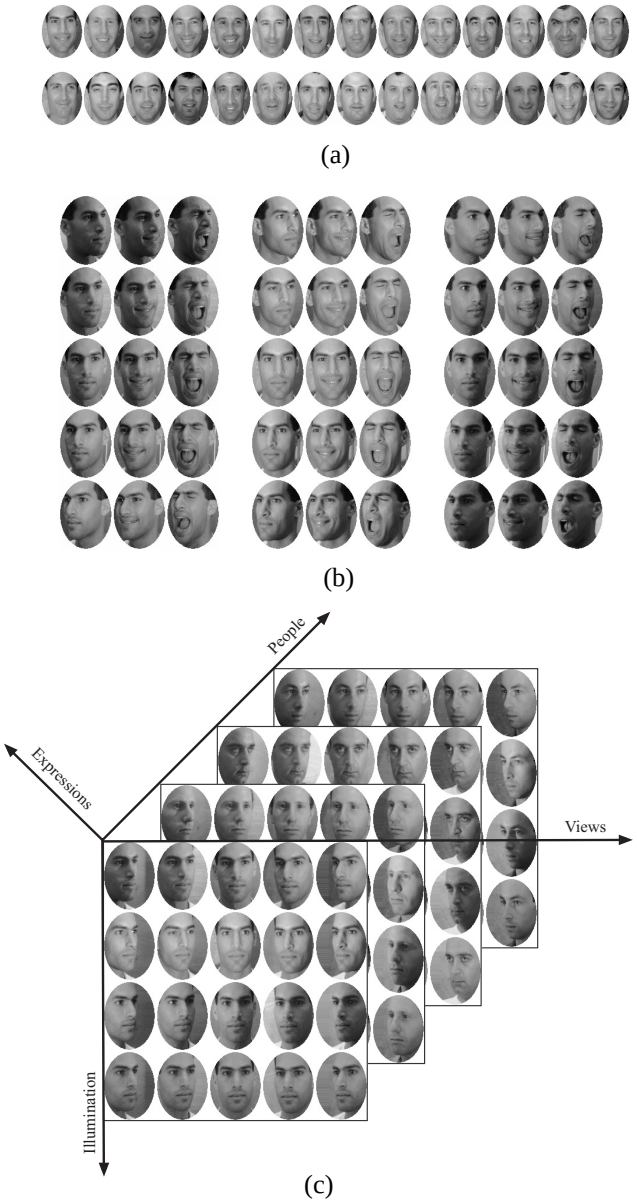
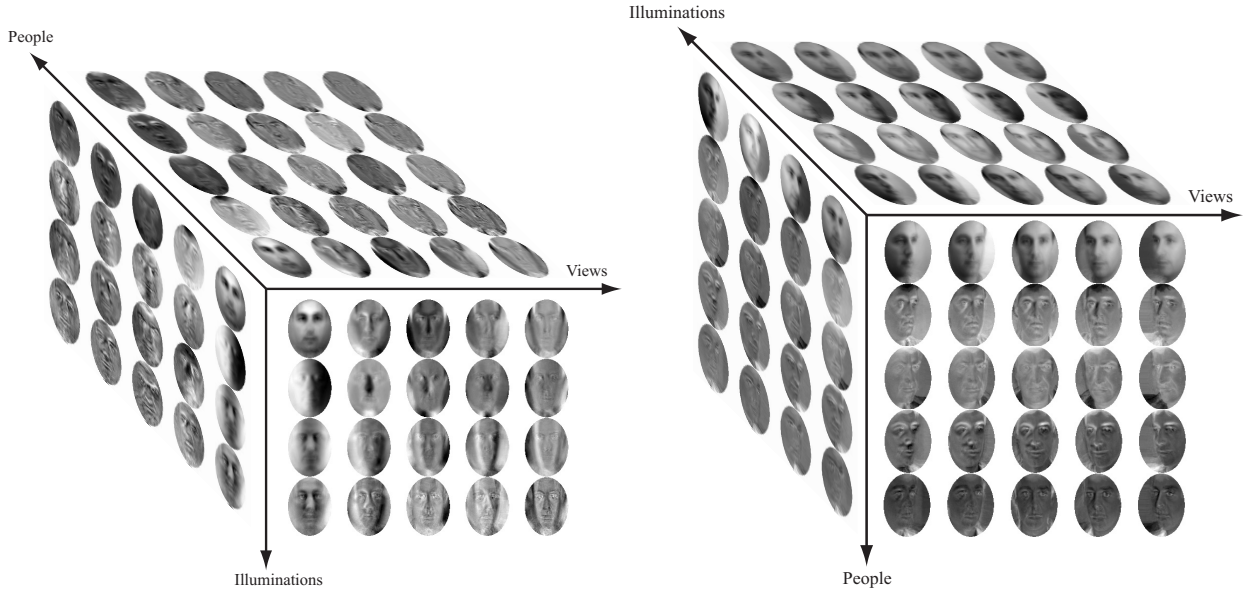


Figure 3: The facial image database (28 subjects, 60 images per subject). (a) The 28 subjects shown in expression 2 (smile), viewpoint 3 (frontal), and illumination 2 (frontal). (b) Part of the image set for subject 1. Left to right, the three panels show images captured in illuminations 1, 2, and 3. Within each panel, images of expressions 1, 2, and 3 (neural, smile, yawn) are shown horizontally while images from viewpoints 1, 2, 3, 4, and 5 are shown vertically. The image of subject 1 in (a) is the image situated at the center of (b). (c) The 5^{th}-order data tensor $\mathcal{D}$ for the image ensemble; only images in expression 1 (neutral) are shown.

Figure 4: $\mathbf{U}_{\text{pixels}}$ contains the PCA eigenvectors (eigenfaces), which are the principal axes of variation across all images.



(a) (b)

Figure 5: (a) A partial visualization of the $28 \times 5 \times 4 \times 3 \times 7943$ TensorFaces representation of $\mathcal{D}$, obtained as $\mathcal{T} = \mathcal{Z} \times_5 \mathbf{U}_{\text{pixels}}$ (only the subtensor of $\mathcal{T}$ associated with expression 1 (neutral) is shown). Note that the mode matrix $\mathbf{U}_{\text{pixels}}$ contains the conventional PCA eigenvectors or "eigenfaces", the first 10 of which are shown in Fig. 4, which are the principal axes of variation across all of the images. (b) A partial visualization of the $28 \times 5 \times 4 \times 3 \times 7943$ tensor $\mathcal{B} = \mathcal{Z} \times_2 \mathbf{U}_{\text{views}} \times_3 \mathbf{U}_{\text{illums}} \times_4 \mathbf{U}_{\text{expres}} \times_5 \mathbf{U}_{\text{pixels}}$ (again, only the subtensor associated with the neutral expression is shown), which defines 60 different bases for each combination of viewpoints, illumination and expressions. These bases have 28 eigenvectors which span the people space. The eigenvectors in any particular row play the same role in each column. The topmost plane depicts the average person, while the eigenvectors in the remaining planes capture the variability across people in the various viewpoint, illumination, and expression combinations.

but it lacks a simple, diagonal structure. The core tensor governs the interaction between the *mode matrices* $\mathbf{U}_1, \ldots, \mathbf{U}_N$. Mode matrix $\mathbf{U}_n$ contains the orthonormal vectors spanning the column space of matrix $\mathbf{D}_{(n)}$ resulting from the *mode-n flattening* of $\mathcal{D}$ (see [10]).

The multilinear analysis of facial image ensembles leads to the TensorFaces representation. We illustrate the technique using a portion of the Weizmann face image database: 28 male subjects photographed in 5 viewpoints, 4 illuminations, and 3 expressions. Using a global rigid optical flow algorithm, we aligned the original 512×352 pixel images relative to one reference image. The images were then decimated by a factor of 3 and cropped as shown in Fig. 3, yielding a total of 7943 pixels per image within the elliptical cropping window.

Our facial image data tensor $\mathcal{D}$ is a $28 \times 5 \times 4 \times 3 \times 7943$ tensor (Fig. 3(c)). Applying multilinear analysis to $\mathcal{D}$, using our N-mode SVD algorithm with $N = 5$, we obtain

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_{\text{people}} \times_2 \mathbf{U}_{\text{views}} \times_3 \mathbf{U}_{\text{illums}} \times_4 \mathbf{U}_{\text{expres}} \times_5 \mathbf{U}_{\text{pixels}}, \quad (2)$$

where the $28 \times 5 \times 3 \times 3 \times 7943$ core tensor $\mathcal{Z}$ governs the interaction between the factors represented in the 5 mode matrices: The 28×28 mode matrix $\mathbf{U}_{\text{people}}$ spans the space of people parameters, the 5×5 mode matrix $\mathbf{U}_{\text{views}}$ spans the space of viewpoint parameters, the 4×4 mode matrix $\mathbf{U}_{\text{illums}}$ spans the space of illumination parameters and the 3×3 mode matrix $\mathbf{U}_{\text{expres}}$ spans the space of expression parameters. The 7943×1680 mode matrix $\mathbf{U}_{\text{pixels}}$ orthonormally spans the space of images. Reference [10] discusses the attractive properties of this analysis, some of which we now summarize.

Multilinear analysis subsumes linear, PCA analysis. As shown in Fig. 4, each column of $\mathbf{U}_{\text{pixels}}$ is an "eigenimage". Since they were computed by performing an SVD of the matrix $\mathbf{D}_{(\text{pixels})}$ obtained as the mode-5 flattened data tensor $\mathcal{D}$, these eigenimages are identical to the conventional eigenfaces. Eigenimages represent the principal axes of variation over *all* the training images. The big advantage of multilinear analysis beyond linear PCA is that TensorFaces explicitly represent how the various factors interact to produce facial images. Tensorfaces are obtained by forming the product $\mathcal{Z} \times_5 \mathbf{U}_{\text{pixels}}$ (Fig. 5(a)).

Multilinear dimensionality reduction generalizes the conventional version associated with linear PCA, truncation of the singular value decomposition (SVD), whose optimality properties are well-known. Unfortunately, *optimal* dimensionality reduction is not straightforward in multilinear analysis. For multilinear dimensionality reduction, we have presented an N-mode orthogonal iteration algorithm that is based the N-mode SVD [11].

The facial image database comprises 60 images per person that vary with viewpoint, illumination, and expression.

PCA represents each person as a set of 60 vector-valued coefficients, one from each image in which the person appears. The length of each PCA coefficient vector is $28 \times 5 \times 4 \times 3 = 1680$. By contrast, multilinear analysis enables us to represent each person, regardless of viewpoint, illumination, and expression, with the same coefficient vector of dimension 28 relative to the bases comprising the $28 \times 5 \times 4 \times 3 \times 7943$ tensor

$$\mathcal{B} = \mathcal{Z} \times_2 \mathbf{U}_{\text{views}} \times_3 \mathbf{U}_{\text{illums}} \times_4 \mathbf{U}_{\text{expres}} \times_5 \mathbf{U}_{\text{pixels}}, \qquad (3)$$

some of which are shown in Fig. 5(b). This many-to-one mapping is useful for face recognition. Each column in the figure is a basis matrix that comprises 28 eigenvectors. In any column, the first eigenvector depicts the average person and the remaining eigenvectors capture the variability over people, for the particular combination of viewpoint, illumination, and expression associated with that column. Each image is represented with a set of coefficient vectors representing the person, viewpoint, illumination and expression factors that generated the image. This is an important distinction that is relevant for facial image recognition.

3.2. Face Recognition Using TensorFaces

We have proposed a recognition method based on multilinear analysis which employs the recognition bases shown in Fig. 5(b) (see [9] for the details). In our preliminary experiments with the Weizmann face image database, TensorFaces yields significantly better recognition rates than PCA (eigenfaces) in scenarios involving the recognition of people imaged in previously unseen viewpoints and illuminations.

In the first experiment, we trained our TensorFaces model on an ensemble comprising images of 23 people, captured from 3 viewpoints ($0, \pm 34$ degrees), with 4 illumination conditions (center, left, right, left+right). We tested our model on other images in this 23 person dataset acquired from 2 *different* viewpoints (± 17 degrees) under the same 4 illumination conditions. In this test scenario, the PCA method recognized the person correctly 61% of the time while TensorFaces recognized the person correctly 80% of the time.

In a second experiment, we trained our TensorFaces model on images of 23 people, 5 viewpoints ($0, \pm 17, \pm 34$ degrees), 3 illuminations (center light, left light, right light) and tested it on the 4^{th} illumination (left+right). PCA yielded a poor recognition rate of 27% while Tensorfaces achieved a recognition rate of 88%.

4. Conclusion

This paper has presented an overview of several model-based and image-based methods that we have developed for facial image synthesis, analysis, and recognition. The two categories of methods are complementary and, in our experience, both are indispensable in tackling the toughest practical problems.

Acknowledgements

The TensorFaces research reviewed herein was funded in part by the Technical Support Working Group (TSWG) through the US Department of Defense's Combating Terrorism Technology Support Program.

References

[1] R. Chellappa, C. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):705–740, May 1995.

[2] P. Ekman, T. Huang, T. Sejnowski, and J. Hager. Final report to NSF of the planning workshop on facial expression understanding. Technical report, National Science Foundation, July 1992. Available on the www.

[3] T. Ishikawa, H. Sera, S. Morishima, and D. Terzopoulos. 3D estimation of facial muscle parameters from 2D marker movements using a neural network. In *Proc. Asian Conf. on Computer Vision*, volume 1352 of *Lecture Notes in Computer Science*, pages 671–678, Berlin, 1998. Springer.

[4] T. Ishikawa, H. Sera, S. Morishima, and D. Terzopoulos. Facial image reconstruction by estimated muscle parameters. In *Proc. Third International Conf. on Automatic Face and Gesture Recognition*, pages 342–347, Nara, Japan, April 1998.

[5] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[6] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *Computer Graphics* Proceedings, Annual Conference Series, Proc. SIGGRAPH '95 (Los Angeles, CA), pages 55–62. ACM SIGGRAPH, August 1995.

[7] C. Pelachaud, N. Badler, and M.-L. Viaud. Final report to NSF of the standards for facial animation workshop. Technical report, National Science Foundation, October 1994. Available on the www.

[8] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579, 1993.

[9] M. Vasilescu and D. Terzopoulos. Multilinear analysis for facial image recognition. In *Proc. Int. Conf. on Pattern Recognition*, pages III–511–514, Quebec City, August 2002.

[10] M. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: TensorFaces. In *Proc. European Conf. on Computer Vision (ECCV 2002)*, pages 447–460, Copenhagen, Denmark, May 2002.

[11] M. Vasilescu and D. Terzopoulos. Multilinear subspace analysis of image ensembles. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages II–93–99, Madison, WI, June 2003.