Joint Modeling of Facial Expression and Shape from Video

T. Tamminen*, J. Kätsyri, M. Frydrych, and J. Lampinen

Laboratory of Computational Engineering,
Helsinki University of Technology,
P.O. Box 9203, 02015 HUT, Finland
toni.tamminen@tkk.fi, katsyri@lce.hut.fi, frydrych@lce.hut.fi,
jouko.lampinen@tkk.fi

Abstract. In this paper, we present a novel model for representing facial feature point tracks during an facial expression. The model is composed of a static shape part and a time-dependent expression part. We learn the model by tracking the points of interest in video recordings of trained actors making different facial expressions. Our results indicate that the proposed sum of two linear models - a person-dependent shape model and a person-independent expression model - approximates the true feature point motion well.

1 Introduction

Human facial expression is a widely studied topic both in computer vision and psychology as a great deal of human communication is carried out through facial expressions, in addition to words. In computer vision, the focus has been on recognition and modeling, while psychologists are interested in both the emotional processes behind facial expressions as well as the brain mechanisms underlying the recognition of emotions from expressions [1]. A most comprehensive system for analyzing facial displays is the Facial Action Coding System (FACS) [2]. It is based on anatomy and has been widely used by psychologists and recently also for automated classification of facial actions. A limitation of FACS is, however, the lack of detailed spatial and temporal information [3]. Improved systems include the FACS+ system [3], the AFA system [4], and many others [5].

Most of the proposed approaches to facial expression modeling are rather holistic in nature, i.e. they model expressions as a whole instead of tracking individual facial feature points. Furthermore, often only the expression is modeled, and no attention is paid to the shape of the face. The combination of these poses a serious problem in some applications such as feature-based object recognition. To deal with the problem, we present a new model for the fiducial feature points of the human face which aims to encompass both the interpersonal facial shape

* Author supported by the Finnish Cultural Foundation, the Jenny and Antti Wihuri Foundation, and the Nokia Foundation.

H. Kalviainen et al. (Eds.): SCIA 2005, LNCS 3540, pp. 151–160, 2005.

variation and the expression-dependent dynamic variation. We aim to represent the sources of variation with orthogonal linear vector bases, which facilitates the analysis and use of the model.

The paper is organized as follows. Section 2 describes our data and our feature tracking system. Section 3 introduces our face model, and Sect. 4 presents analysis of the model and some reconstruction results. Section 5 concludes.

2 Data Acquisition and Feature Tracking

2.1 The Data

Facial expressions were recorded from actors trained to express certain prototypical emotional facial expressions. The recordings included seven facial expressions related to basic emotions [6] (two different happiness recordings), two facial expressions related to blends of basic emotions and one emotionally meaningless facial expression. The facial expression prototypes (Table 1) were based on existing descriptive literature [2] [7] and defined for FACS by a certified FACS coder.

The recordings were made from 6 actor students from the Theatre Academy of Finland (3 men and 3 women, age range 23-32 years); hence, there were 60 video streams in total. The actors were asked both to express the given facial configuration exactly and to experience the required emotion. The actors

Table 1. Facial Expression Prototypes

Facial expression	FACS Action units	Facial expression	FACS Action units
Anger	4+5+7+24	Sadness	1+4+7+15+17
Disgust	9+10+17	Surprise	1+2+5+25+26
Fear	1+2+4+5+7+20+25	Happiness + surprise	1+2+5+6+12+25+26
Happiness (mouth open)	6+12+25	Happiness + disgust	6+9+10+12+17
Happiness (mouth closed)	6+12	Mouth opening	25+26

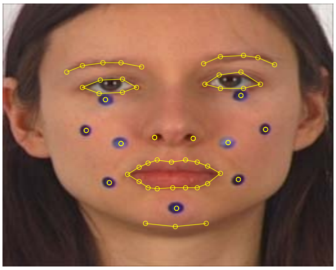

Fig. 1. A sample feature graph, with the added dark markers showing. The light dots mark the tracked features

practised the facial expressions individually for approximately 5-10 hours. One practise recording was carried out with the possibility for feedback before the actual recording session.

The recordings contained short (1-2s.) video sequences showing the change from neutral to the target state. Nine markers were placed on perceptually meaningful locations (Fig. 1) to ease the tracking of facial changes unrelated to clear facial features. The recording setup included 2 professional photographing lamps (Elinchrom scanlite 1000) and a digital camcorder (Sony DSR-PD100AP). The recordings were made at 25 (interlaced) frames per second (fps) with a resolution of 572*726 pixels. To reduce computational cost and memory usage, the videos were clipped to include only the facial area and resized to 256*256 pixels.

2.2 Feature Tracking

The KLT tracker and its derivatives are used widely in visual feature tracking [8] [9] [10]. However, we decided to test the possibilities of an automated tracker based on Gabor filters [11] and Bayesian inference [12] as an extension of our static object matching system [15]. A similar approach (without the Bayesian context) has previously been presented by Mckenna et al. [13].

To reduce clutter and to make the features more distinctive, each image $\mathbf{I}^t$ in a video sequence (with time steps t) is first transformed into feature space, $\mathbf{I}^t \mapsto \mathbf{T}^t$, by filtering it with a Gabor filter bank with 3 frequencies and 6 orientations. All computations are then performed using the transformed images $\mathbf{T}^t$. The face is represented as a planar graph containing n nodes (Fig. 1) with coordinates $\mathbf{X}^t = \{x_1^t, ..., x_n^t\}$. Each node i has an associated feature vector $\mathbf{g}_i^t$, which is formed by stacking the responses of the Gabor filter bank as a vector.

The features are tracked by finding, at each time step, the maximum a posteriori estimate of the location of each feature around its previous location. That is, we compute the posterior density of each feature in some search area A_i given the transformed image, the corresponding feature vector $\mathbf{g}_i^t$ and the other feature locations $\mathbf{x}_{\setminus i}^t$, and maximize it:

$$\max_{x_i^t \in A_i} p(x_i^t | \mathbf{T}^t, \mathbf{g}_i^t, \mathbf{x}_{\setminus i}^t) \propto p(\mathbf{T}^t | x_i^t, \mathbf{g}_i^t) p(x_i^t | \mathbf{x}_{\setminus i}^t), \tag{1}$$

where we have used Bayes's formula to write the posterior probability as the product of the likelihood and prior parts. The likelihood measures the probability of observing the image given a feature configuration, while the prior gives the distribution of the feature location given the locations of the other features.

We can not measure the probability of observing an image directly, since we do not have a comprehensive image model. Hence, we approximate the likelihood by computing the similarity between the stored feature vectors $\mathbf{g}$ and the perceived image $\mathbf{T}$. We use the criterion presented by Wiskott et al. [14] to obtain the similarities. As the prior we use a simple model in which the features are allowed independent Gaussian variations from a known mean shape $\mathbf{r}^t$ [15]:

$$p(x_i^t | \mathbf{x}_{\setminus i}^t) = \mathrm{N}(f(\mathbf{x}_{\setminus i}^t, \mathbf{r}^t), \sigma^2), \tag{2}$$

where f is a function that translates and scales the mean shape to correspond to the current graph, and σ^2 is the variance of the Gaussian, which was set to some suitable value so that the tracker would function well (with 256 × 256 images, we used $\sigma = 5$).

As the video sequence progresses, both the features $\mathbf{g}$ and the mean shape $\mathbf{r}$ change. To adapt the tracker to this, at each time step we change $\mathbf{g}$ and $\mathbf{r}$ according to the newly obtained values:

$$\mathbf{g}^{t+1} = \alpha_g \mathbf{g}^t + (1 - \alpha_g)\mathbf{g}^1 \qquad (3)$$

$$\mathbf{r}^{t+1} = \alpha_r \mathbf{r}^t + (1 - \alpha_r)\mathbf{r}^1, \qquad (4)$$

where α_g and α_r are parameters controlling the extent of the adaptation. Using $\mathbf{g}^1$ and $\mathbf{r}^1$ as the baseline values reduces the probability of the tracker adapting to track a completely spurious feature, as the effect of the original Gabor jets and mean shape never disappears completely.

The initial feature locations $\mathbf{X}^1$ and Gabor jets $\mathbf{g}^1$ are obtained by manually annotating the features on the first image of one video sequence and then using the image and the annotations as training data for matching the features in the first images of other sequences (for details of the matching, see [15]). The mean shape $\mathbf{r}^1$ is taken to be equal to $\mathbf{x}^1$.

The performance of the tracker was varying. In some streams it tracked the features perfectly, in some streams there were considerable errors. The tracking could be improved in numerous ways such as including a systematic model for the motion of the features or designing a more sophisticated adaptation scheme. However, since the tracking was not the main object of interest in this paper, the improvements were left to a further study.

3 Face Model

In our model, our aim is to find separate orthogonal bases for representing variations due to face shape and facial expression. A similar approach has been proposed by Abboud and Davoine [16]; however, they do their modeling in the AAM framework [17] and model only the start- and endpoints of expressions, whereas we are interested in the the whole track of the fiducial feature points during an expression.

To model the dynamics of the expression, we include the time correlations of the feature point tracks into our expression model, that is, the expressions are described by vectors of length $n \times n_t$, where n_t is the number of time steps. We assume that the tracks $\mathbf{X} = \{\mathbf{X}^1, ..., \mathbf{X}^{t_f}\}$ can be represented as the sum of two linear models: a person-dependent shape model and a person-independent expression model so that

$$\mathbf{X} = \mathbf{1} \otimes (\mathbf{m} + \mathbf{S}\beta_{person}) + \mathbf{E}\beta_{expression} + \epsilon, \qquad (5)$$

where $\mathbf{m}$ is the mean shape, $\mathbf{S}$ is the matrix of the base vectors of the shape space, $\mathbf{E}$ is the matrix of the base vectors of the expression space, β_{person} is the person-dependent vector of coordinates in the shape space, $\beta_{expression}$ is the expression-dependent vector of coordinates in the expression space, $\mathbf{1}$ is a vector of ones, $\otimes$ is the Kronecker product, and ϵ is Gaussian noise. Note that the Kronecker product is required to make the the computation of the sum possible, as the shape and expression vectors are of different lengths. At time step t the graph is

$$\mathbf{X}^t = \mathbf{m} + \mathbf{S}\beta_{person} + \mathbf{E}^t\beta_{expression} + \epsilon^t, \tag{6}$$

where $\mathbf{E}^t$ contains the elements of the expression base vectors that apply to time step t.

To estimate the base vectors of the shape and expression spaces, we need to separate the shape and expression effects. This is done in two phases:

1. Estimate the mean shape and the shape base vectors via PCA [18] from the initial feature graphs $\mathbf{X}^1$. We assume that the video streams start from a neutral expression, that is, $\mathbf{E}^1 = 0$.
2. To remove the effect of the shape from subsequent images in the stream, subtract the projection of the initial graph onto the shape base $\mathbf{S}\mathbf{S}^T\mathbf{X}^1$ from the subsequent graphs. Then stack the graphs as vectors and perform PCA to obtain the expression base vectors.

Note that in phase 2, the PCA is perfomed on the correlation matrix of the vectors, that is, we do not subtract a "mean expression" from the graphs.

The model can be described also in a slighly different way as the sum of two Gaussian distributions:

$$p(\mathbf{X}) = \mathbf{1} \otimes \mathrm{N}(\mathbf{m}, \Sigma_{shape}) + \mathrm{N}(0, \Sigma_{expression}), \tag{7}$$

where Σ_{shape} is the covariance matrix of the shape distribution and $\Sigma_{expression}$ the correlation matrix of the expression distribution (with $\mathbf{S}\mathbf{S}^T\mathbf{X}^1$ removed). The eigenvectors of these matrices are the base vectors mentioned above.

In practice we need to normalize our tracking results before they can be used to learn the model parameters. First we translate and scale the graphs so that their mean locations and their scale factors are the same. We define the scale factor as

$$s = \sqrt{0.5\sigma_x^2 + 0.5\sigma_y^2}, \tag{8}$$

where σ_x and σ_y are the standard deviations of the graph x- and y-coordinates. Then, to make the model symmetrical, we insert a mirrored replicate graph for every measured graph in the data. Finally, the lengths of the tracks are normalized by selecting a common frame number (larger than the length of the longest video sequence) and interpolating the tracks as necessary so that their lengths match.

4 Analysis and Reconstruction

To analyze the model and assess its capabilities, we performed a set of reconstruction-related tests. The shape and expression bases were computed using the measured tracking results and the principal components were inspected visually. The first two expression principal components are illustrated in Fig. 2. We then projected the measured tracks onto the obtained bases and analyzed the coordinates to see whether our separability assumption (person-dependent shape, person-independent expression) held. Some projection coordinate plots are shown in Fig. 3 and Fig. 4. It would seem that the separability assumption holds: the shape space coordinates remain in most cases approximately equal for the same person, while the expression space coordinates are similar for the same expression.

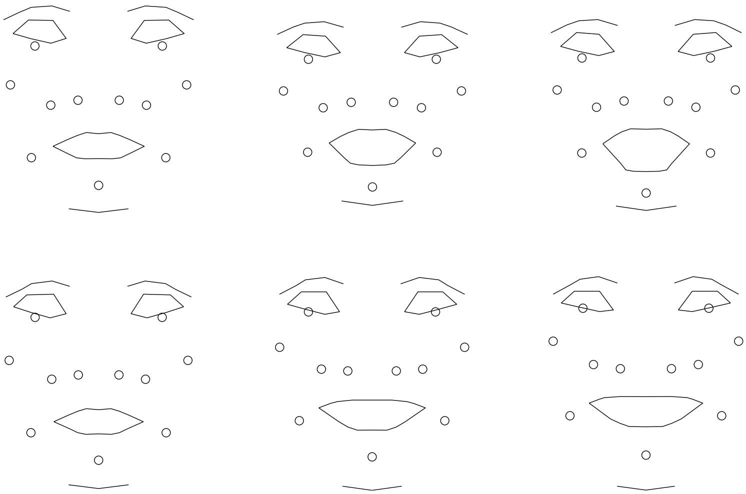

Fig. 2. The first two expression principal components. The components are shown at time steps $t = 1$, $t = 1/2t_f$ and $t = t_f$. The first component (row 1) is mainly related to opening of the mouth, while the second component (row 2) seems to be a smile

The actual reconstruction was done by projecting the measured tracks into the shape and expression spaces and then back to the original track space to obtain the reconstructed tracks $\mathbf{X}^*$,

$$\mathbf{X}^* = \mathbf{1} \otimes (\mathbf{m} + \mathbf{S}\mathbf{S}^T\mathbf{X}^1) + \mathbf{E}\mathbf{E}^T\mathbf{X}. \tag{9}$$

We used 15 principal components for the shape space and 6 components for the expression space, which in both cases amounted to ca. 99% of the total variance. The original and reconstructed tracks were compared both visually and numerically. Two sample reconstructions are shown in Fig. 5, and Table 2

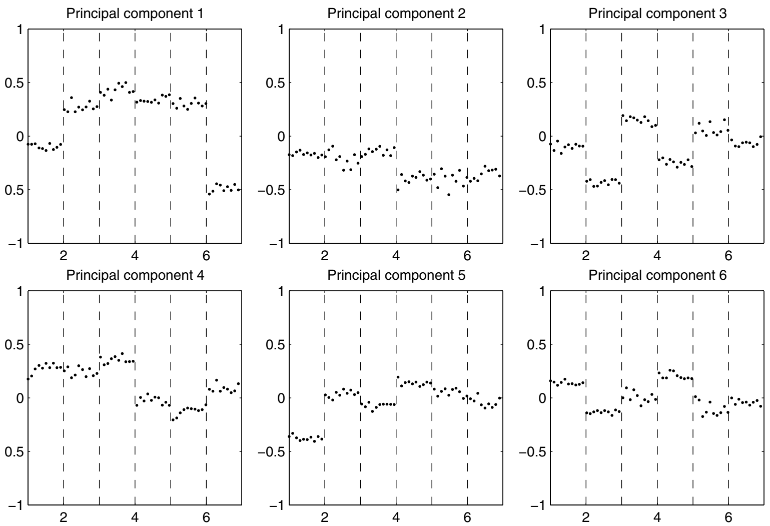

Fig. 3. First six shape space coordinates for the 60 initial graphs $\mathbf{X}^1$. The x-axis is the person index from 1 to 6. Each image corresponds to a single principal component with 10 coordinate instances for each person. The dashed lines indicate change of person. In most cases, the persons are clearly distinct from one another, and the coordinates are similar for the same person

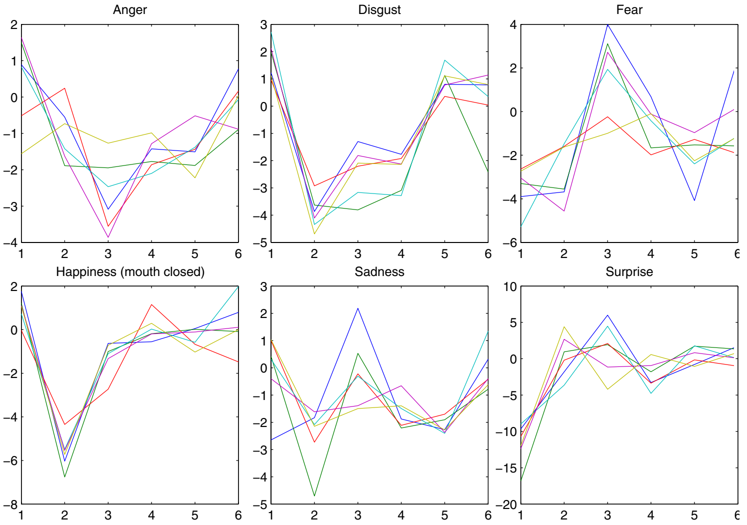

Fig. 4. First six expression space coordinates for the six basic expressions. The x-axis is principal component index. Each line corresponds to a single expression instance. The expressions are similar to each other across persons, although there are differences, too. For example, the coordinates for the expressions of happiness show more similarity than the expressions of fear. The similar situation is encountered in everyday life - expressions of happiness are much more alike than expressions of fear

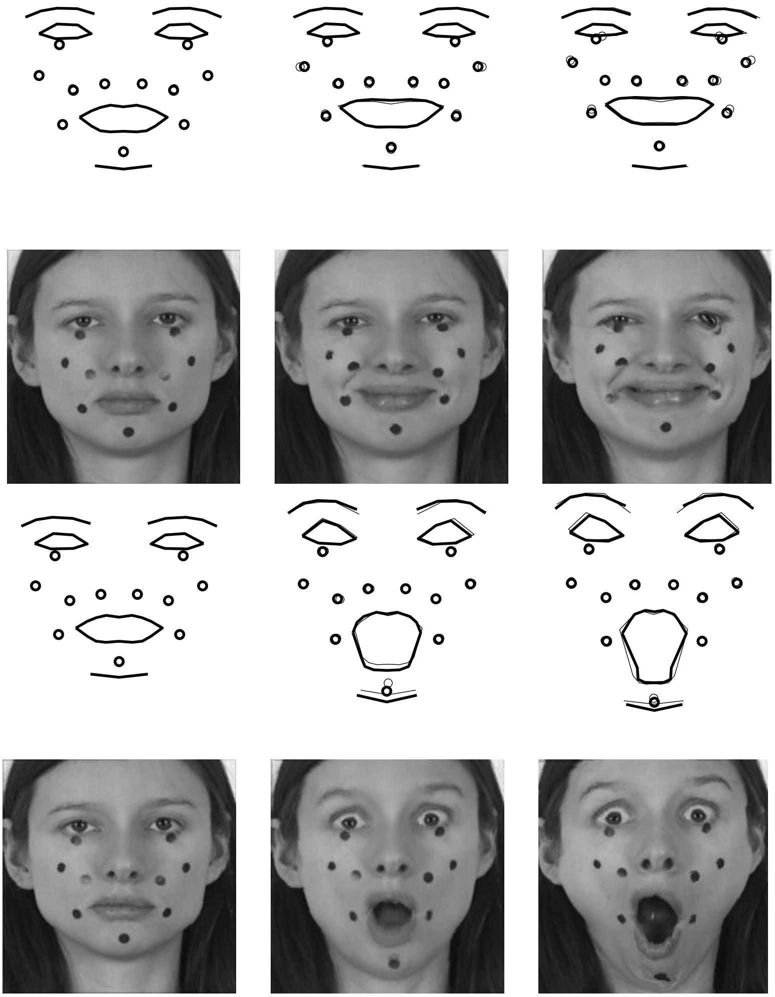

Fig. 5. Reconstruction results for the "happiness (mouth closed)" (upper two rows) and "surprise" (lower two rows) expressions. The depicted time steps are $t = 1$, $t = 1/2t_f$ and $t = t_f$. The thinner graphs show the original data and the thicker graphs the reconstructed expressions, while the images show the results of morphing the video frame corresponding to the time step according to the reconstructed graph. The expressions are clearly recognizable, and there are few distortions

contains mean reconstruction errors per unit of scale as defined by the scale factor (8) (for the unscaled size 256 × 256 training data the scale was around 50).

The reconstruction results are rather promising: visually, the reconstructed expressions are easily recognizable and contain little distortion, and the numerical errors are low (for the original data, the mean error is below 2 pixels for most cases).

Table 2. Mean Reconstruction Error per Unit of Scale

Expression	$t = 1$	$t = 1/2t_f$	$t = t_f$	$t = \{1...t_f\}$
Anger	0.0070	0.0267	0.0353	0.0214
Disgust	0.0071	0.0225	0.0296	0.0198
Fear	0.0082	0.0274	0.0353	0.0221
Happiness (mouth open)	0.0069	0.0250	0.0336	0.0208
Happiness (mouth closed)	0.0061	0.0246	0.0356	0.0212
Sadness	0.0073	0.0240	0.0311	0.0206
Surprise	0.0071	0.0265	0.0322	0.0229
Happiness + surprise	0.0072	0.0337	0.0411	0.0251
Happiness + disgust	0.0078	0.0282	0.0385	0.0246
Mouth opening	0.0063	0.0221	0.0258	0.0174
All expressions	0.0071	0.0261	0.0338	0.0216

5 Conclusion

We have presented a novel model for the representation of fiducial feature points on the human face. The model is a sum of two linear submodels: a person-dependent shape model and a person-independent expression model. The parameters of the model are learned from video data of trained actors making specified expressions. Our reconstruction results imply that the proposed separation of the facial graph as orthogonal shape and expression parts is feasible.

The model presented here is trained only on frontal facial images, and can not handle large pose variations. With 3D data it should be straightforward to extend the model to accommodate these. Also, there is considerable intrapersonal variation in facial expressions with regard to their strength and speed, whereas the current model assumes that expression durations and speeds are the same. This problem has to be addressed in further research.

The model has several practical applications. In its probabilistic form (7) the model can be used directly as a prior in expression-dependent Bayesian object matching [15]. Furthermore, in the future we will work on implementing the expressions on a Talking Head model [19]. The proposed model includes the dynamics of the expressions, and hence should be an improvement over the previously used expression model. Another interesting research topic is to compare the obtained expression principal components (Fig. 2) and the FACS action units to see whether there is any systematic correspondence.

References

1. R. Adolphs, Recognizing emotion from facial expressions: psychological and neurological mechanisms, *Behavioral and Cognitive Neuroscience Reviews*, vol. 1, no. 1, 2002, pp. 21-62.
2. P. Ekman, W. Friesen, and J. Hager, *Facial Action Coding System*, Consulting Psychologists Press, 1978.

3. I.A. Essa and A.P. Pentland, Coding, analysis, interpretation and recognition of facial expressions, *IEEE TPAMI*, vol. 19, no. 7, 1997, pp. 757-763.
4. Y.-l. Tian, T. Kanade, and J.F. Cohn, Recognizing action units for facial expression analysis, *IEEE TPAMI*, vol. 23, no. 2, 2001, pp. 97-115.
5. G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, Classifying facial actions, *IEEE TPAMI*, vol. 21, no. 10, 1999, pp. 974-989.
6. P. Ekman, Expression and the nature of emotion, In: K. Scherer and P. Ekman, editors, *Approaches to Emotion*, Lawrence Erlbaum, 1984.
7. P. Ekman and W. Friesen, *Unmasking the Face. A Guide to Recognizing Emotions from Facial Expressions*, Consulting Psychologists Press, 1975.
8. B.D. Lucas and T. Kanade, An iterative image registration technique with an application to stereo vision, In: *Proc. Imaging Understanding Workshop*, 1981, pp. 121-130.
9. C. Tomasi and T. Kanade, Detection and tracking of feature points, *Carnegie Mellon University Technical Report CMU-CS-91-132*, 1991.
10. F. Bourel, C.C. Chibelushi, and A.A. Low, Robust Facial Feature Tracking, In: *Proc. BMVC 2000*, 2000.
11. J.G. Daugman, Complete discrete 2-d Gabor transforms by neural networks for image analysis and compression, *IEEE TASSP*, vol. 36, no. 7, 1988, pp. 1169-1179.
12. A. Gelman, J.B. Carlin, H.S. Stern, and D.R. Rubin, *Bayesian Data Analysis*, 2nd edition, Chapman & Hall, 2004.
13. S.J. McKenna, S. Gong, R.P. Wurtz, J. Tanner, and D. Banin, Tracking facial feature points with Gabor wavelets and shape models, In: *Proc. 1st International Conference on Audio- and Video-based Biometric Person Authentication*, 1997.
14. L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, Face Recognition by Elastic Bunch Graph Matching, In: L.C. Jain, U. Halici, I. Hayashi, S.B. and Lee, editors, *Intelligent Biometric Techniques in Fingerprint and Face Recognition*, CRC Press, 1999.
15. T. Tamminen and J. Lampinen, Bayesian object matching with hierarchical priors and Markov chain Monte Carlo, In: J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West, editors, *Bayesian Statistics 7*, Oxford University Press, 2003.
16. B. Abboud and F. Davoine, Appearance factorization for facial expression analysis, In: *Proc. BMVC 2004*, 2004, pp. 507-516.
17. T.F. Cootes, G.J. Edwards, and C.J. Taylor, Active appearance models, *IEEE TPAMI*, vol. 23, no. 6, 2001, pp. 681-685.
18. C. Chatfield and A.J. Collins, *Introduction to Multivariate Analysis*, Chapman & Hall, 1995.
19. M. Frydrych, J. Kätsyri, M. Dobšík, and M. Sams, Toolkit for animation of Finnish talking head, In: *Proc. AVSP 2003*, 2003, pp. 199-204.