A Flexible Object Model for Recognising and Synthesising Facial Expressions

Andreas Tewes, Rolf P. Würtz and Christoph von der Malsburg

Ruhr-Universität, Institut für Neuroinformatik, D–44780 Bochum, Germany

Abstract. We here introduce the Flexible Object Model to represent objects with structured deformation, such as the human face under variable expression. The model represents object shape and texture separately and extracts a data parameterisation autonomously from image sequences after initialisation by a single hand-labeled model graph. We apply the model to the representation, recognition and reconstruction of nine different facial expressions. After training, the model is capable of automatically finding facial landmarks, extracting deformation parameters and reconstructing faces in any of the learned expressions.

1 Introduction

Elastic matching of graphs labeled with Gabor wavelet features (EGM) [1] has proved a very successful basis for invariant object recognition, even when spatial deformation is involved as with face recognition under small changes in pose or expression. According to that concept, variation due to position, scale and in-plane orientation can be dealt with precisely, but intrinsic image deformations are not actively modeled and can only passively be followed. This leads to limited discriminatory power during recognition and precludes the possibility to reconstruct images from model data. Facial image deformations due to pose or expression are highly structured and should be represented by a parameterised model. To this end we have developed a Flexible Object Model (FOM). It continues to use elastic graphs to represent objects in individual images but parameterises these graphs, treating them as functions of pose and expression parameters. In this paper we present the FOM in general and apply it in chapter 3 to the description of the human face under nine different expressions. We demonstrate the power of the model by matching and reconstructing faces in a person-independent way. We conclude by discussing possible applications, among them improved facial recognition under variable expression.

2 The Flexible Object Model

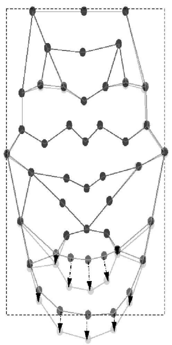
The FOM, using Gabor wavelet-labeled graphs as fundamental data structure, distinguishes object shape (represented by the spatial arrangement of landmarks) from texture (represented by Gabor jets attached to the landmarks). While deformation of shape is described in a parameterised way relative to a reference model, the interrelationship between shape deformation and texture is characterised using a linear function mapping the former onto the latter. The FOM therefore also includes mappings between shape deformation and texture, an idea developed earlier in our lab [2, 3]. Both the variations (of shape and texture) and the mappings between them are extracted by statistical procedures from video frame sequences for one or several persons performing different facial gestures. Compared to the concept of *Active Appearance Models*, which describes shape and texture variations using either one common set of parameters or one set for each [4], in the context of FOM only the shape variation is learned in a parameterised way while the texture is assumed to be fully determined by a given shape and map. Finally also the matching process, which uses the concept of EGM and is described in section 4 in detail, differs from the *Fitting* process in the context of AAMs.

2.1 Data Collection

We used sample material collected by Hai Hong [3]. The sequences were taken under fairly controlled lighting conditions and in frontal pose. In each sequence the subject performs one of a number of facial gestures, starting and ending with neutral expression. The gestures were selected for ease of performance (shunning the difficulty of expressing emotional states) and attempting to cover the whole range of facial movements. In this study we have used only a subset of 9 of the 23 gestures originally collected [3] for each person.

We initialise the process of extracting model graphs from the frames of a sequence by manually locating the nodes of the standard graph over facial landmarks in the first frame. The system then automatically tracks these nodes from frame to frame with a method based on the phases of Gabor wavelets [5]. The link structure of the graphs is kept constant throughout. For the sake of scale invariance, the size of the reference graph is noted in terms of a bounding box and node displacement in x- and y-direction is measured relative to the width and height of that box, respectively. To encode local image texture, responses of several Gabor kernels differing in scale and orientation [1] were extracted at the landmark positions in each frame. We treat a set of 300 Gabor responses (real and imaginary part, 15 orientations [$\phi_{min} = 0$, $\phi_{max} = \frac{14}{15}\pi$] and 10 scales [$k_{min} = 2^{-\frac{11}{2}}\pi$, $k_{max} = \frac{1}{2}\pi$]) as one real-valued vector, called Gabor jet.

For each frame, the normalised shift vectors of landmarks relative to the first frame as well as the Gabor jets at the node positions are noted. They together form the raw input data, see (1). The models of section 3 were created for individuals, the models in sections 4 and 5 were formed using video sequences for several persons. Figure 1, left side, shows two facial graphs superimposed on each other. The graph with black nodes represents the reference shape (first frame) while the one with grey nodes belongs to a deformation (which in this case obviously only affects mouth and chin).



$$\Longleftrightarrow \quad \begin{pmatrix} \dfrac{\Delta x_1}{w} \\ \dfrac{\Delta y_1}{h} \\ \vdots \\ \dfrac{\Delta x_N}{w} \\ \dfrac{\Delta y_N}{h} \end{pmatrix}$$

Fig. 1. Deformation from reference shape normalised by the width w and height h of the bounding box. N denotes the number of nodes while Δx_i and Δy_i indicates the displacement of node i along x- and y-direction.

2.2 Model Formation

To construct the FOM as a parameterised model of graph deformation, the raw data extracted from several video sequences are merged using *Principal Component Analysis* (PCA) [6] and a *Neural Gas* (NG) [7]. While the latter is suitable for forming sparse representations of the extracted deformations and for classification purposes, PCA is important for data compression and is particularly interesting for interpolating and extrapolating the deformations present in the samples. By this, different deformations which do not not occur simultaneously in the sample sequences can be superimposed, as illustrated in figure 4. In addition we are working with *Principal Curves* [8] to describe smooth transitions, although we don't elaborate on that here.

To represent our raw data we use the following notation. If the number of video frames and raw graphs is M we form the matrices

$$\underline{\underline{D}} := (\boldsymbol{d_1} \ldots \boldsymbol{d_M}) \, ; \quad \underline{\underline{F^i}} := (\boldsymbol{j^i_1} \cdots \boldsymbol{j^i_M}) \, , \tag{1}$$

where the column vector $\boldsymbol{d}$ denotes the deformation as introduced in figure 1 and the column vector $\boldsymbol{j^i}$ indicates the feature vector belonging to the node with index i. Using PCA, we can now construct the following quantities

$$< \boldsymbol{D} > \equiv \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{d}_m \tag{2}$$

$$< \boldsymbol{F^i} > \equiv \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{j^i_m} \tag{3}$$

$$\underline{\underline{P}} := (\boldsymbol{P_1} \ldots \boldsymbol{P_L}) \equiv \text{Principal Deformations} \tag{4}$$

$$\underline{\underline{Q}}^i := (\boldsymbol{Q^i_1} \ldots \boldsymbol{Q^i_K}) \equiv \text{Principal Features at node } i \tag{5}$$

$$\underline{\underline{\tilde{D}}} := \underline{\underline{D}} - < \boldsymbol{D} > \underbrace{(1 \ldots 1)}_{M \text{ times}} \equiv \text{Mean-Free Deformations} \qquad (6)$$

$$\underline{\underline{\tilde{F}}}^{i} := \underline{\underline{F}}^{i} - < \boldsymbol{F^i} > \underbrace{(1 \ldots 1)}_{M \text{ times}} \equiv \text{Mean-Free Features of node } i \qquad (7)$$

where all vectors are taken as column vectors. To reduce the data dimensionality we use only the first **L** principal components to describe graph deformation and the first **K** principal components for the Gabor jets, respectively. Throughout this paper we have set $L = 7$ and $K = 20$, values that proved sufficient to reproduce the original data with little error.

Shape deformation is always accompanied by changing texture. We make the simple assumption of a linear mapping between the shape deformation and the feature vectors (or rather their mean-free versions), and see that assumption justified by our numerical results, see chapter 3. Using the matrices $\underline{\underline{A}}^i$ (one matrix per node) we can express and estimate this relationship as follows,

$$\underline{\underline{A}}^i \underline{\underline{P}}^T \underline{\underline{\tilde{D}}} \overset{!}{=} (\underline{\underline{Q}}^i)^T \underline{\underline{\tilde{F}}}^i \quad \Rightarrow \quad \underline{\underline{A}}^i \approx (\underline{\underline{Q}}^i)^T \underline{\underline{\tilde{F}}}^i \left(\underline{\underline{P}}^T \underline{\underline{\tilde{D}}} \right)^+ , \qquad (8)$$

where + indicates the Moore-Penrose inverse [9] of the term in brackets. By using homogeneous coordinates it is possible to squeeze all necessary operations into one matrix that maps a given deformation immediately onto the feature vector. This is important because it accelerates the computation and therefore makes it more suitable for the matching tasks introduced in chapter 4.

3 Flexible Model for synthesising Facial Expressions

In this section we demonstrate the ability of the FOM to synthesise images of varying facial expression. To this end we have created a person-specific FOM, using as data nine video sequences with nine different facial expressions (each containing between 30 and 70 frames). Sample frames are shown in figure 2.

Figure 3 shows three sample frames, taken from the same sequence, with tracked landmarks.

After collecting the data from all nine sequences, we perform the PCA of steps (4) and (5), and estimate the shape-to-texture mappings according to (8). To demonstrate the resulting FOM we chose two of the principal components, added them with variable amplitude to the mean deformation (which is near to the neutral expression) and show in figure 4 reconstructions of the resulting data models. Reconstructions were obtained by the method of [10]. In the bottom row of the figure the PC amplitude runs from one negative standard deviation on the left through zero in the middle to one positive standard deviation on the right. The middle column shows the effect of another PC for positive amplitudes. Three of the gestures shown in figure 2 can be recognised among the reconstructions in the middle columns and bottom row. The diagonals of the figure were formed by superposition of the two PCs and show gestures not present in the input sequences, demonstrating the extrapolation alluded to above.

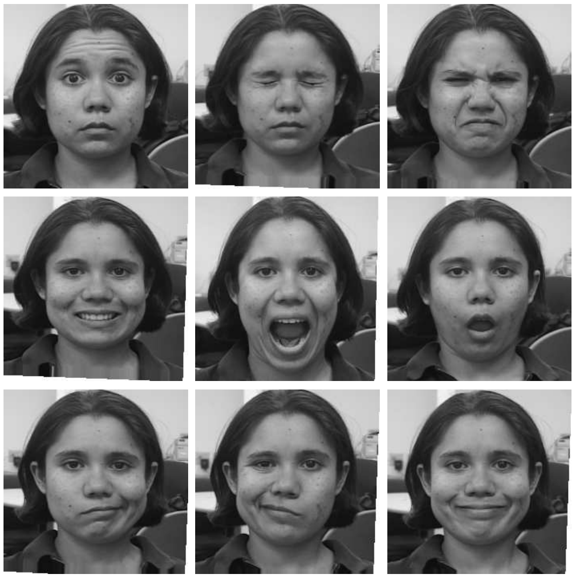


Fig. 2. Facial Gestures shown at maximal extent.

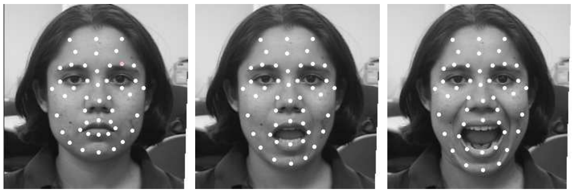


Fig. 3. Autonomously Tracked Landmarks within a gesture sequence.

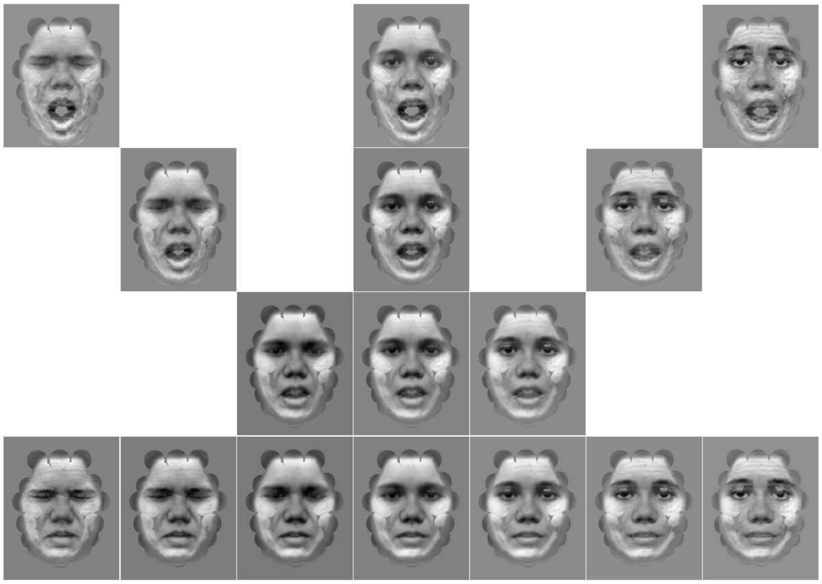


Fig. 4. Synthesised facial expressions using the first (shown vertically) and fourth (shown horizontally) Principal Deformation as well as superpositions.

In the next section we will need a discrete set of "canonical" facial deformations. To this end we use a Neural Gas [7] for clustering and apply the following procedure. From each frame we obtain a shape deformation vector $\mathbf{d}$. This we project into the subspace of the first $L = 7$ principal components. These shape vectors are clustered by a neural gas of 9 neurons, each neuron corresponding to a cluster. Figure 5 shows the deformed face graphs for the 9 neurons or clusters. From each neuron's deformation $\mathbf{d}$ we obtain Gabor jets by applying the matrices $\underline{\underline{A}}^i$ and reconstruct a facial image, shown for the 9 clusters or canonical gestures in figure 6.

4 Landmark Finding

Landmark finding, that is, the establishment of precise point correspondences between facial images and generic facial models, is a critical step for facial processing. It is difficult to achieve, especially in the presence of facial deformation. *Passive* mechanisms, such as classical elastic graph matching [1] have to be permissive in terms of deviations of image texture and shape relative to the model and thus lose reliability quickly beyond small deformations. The problem can only be solved by *actively* modeling the changes in texture and shape observed in images. For this purpose we here employ a FOM. For greater robustness we have trained it on four different persons (where we used the total number of se-

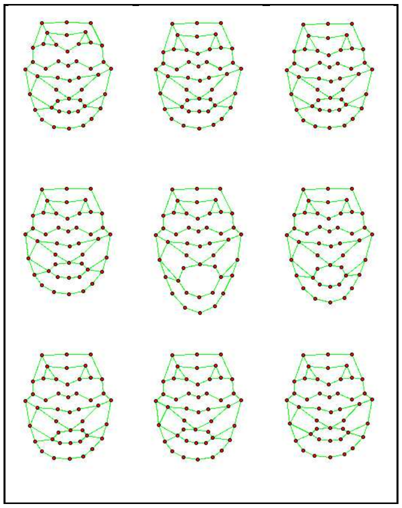


Fig. 5. Shape deformations as per Neural Gas. Expressions are shown corresponding to figure. 2

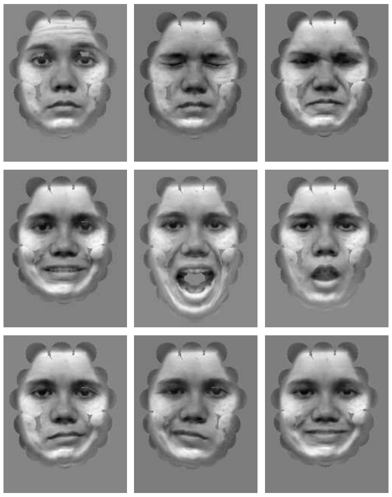


Fig. 6. Synthesised images using shape deformations as shown in figure 5.

quences collected from all persons while each person contributes a data amount as described in section 3).

Our test images display facial gestures of persons not contained in the data set used for training the FOM. We first find the face to some precision with the help of a bunch graph [1] constructed out of frontal neutral-expression images for six persons (again different from the test persons). After suppressing the background of the image outside the convex hull of the bunch graph by Gaussian smoothing we replace the bunch graph by the graph of the FOM and improve the match by separate scale moves in vertical and horizontal directions using the reference shape. Starting from this reference graph, we now apply the nine "canonical" gesture deformations trained by the methods of the last section on four persons (each with the amplitude represented by the trained neurons) and pick the best-matching gesture. Figure 7 shows examples for six different facial expressions. The first and third column show test images with suppressed background and superimposed best-matching graph, each image accompanied on its right by a reconstruction from the 4-person FOM in the best-matching expression.

In addition to accurate landmark finding in spite of image deformation the system can be used to identify the gesture displayed in the image. Using several persons to construct the FOM increased the robustness of the model for person-independent matching (just as the bunch graph increases the robustness of face finding), and in addition handled personal differences in the reference persons' performance of gestures (although the gestures were originally selected for ease of performance [3]).

5 Correction of Facial Expression

An important application of our FOM will be face recognition. Even for collaborating subjects, variation in facial expression cannot be totally avoided and passive methods of dealing with it are compromising accuracy. What is required is active modeling of the effect of expression change so that the test image's expression can be adjusted to that of the gallery entry (or vice versa). After that step, standard recognition tools can be used. We here show in exploratory experiments that our FOM is a viable basis for this operation.

Without loss of generality we assume that images in the gallery are of neutral expression. Using a FOM, trained as described in the previous two sections on data for several (4) persons, we first recognise the expression in the test image by selecting the best-matching canonical expression (including neutral). After landmark finding, feature vectors are extracted from the test image and the face is transformed with the help of the FOM into neutral expression by applying the reference shape and replacing only those jets which are *significantly* deformed with the corresponding jets of the neutralised FOM. By keeping the jets which belong to landmarks hardly deformed as much as possible of the subject's identity should be preserved. An example of this approach is shown in figure 8. The thus synthesised model is compared with the one stored in the database. A similar approach can be applied to changing head pose.

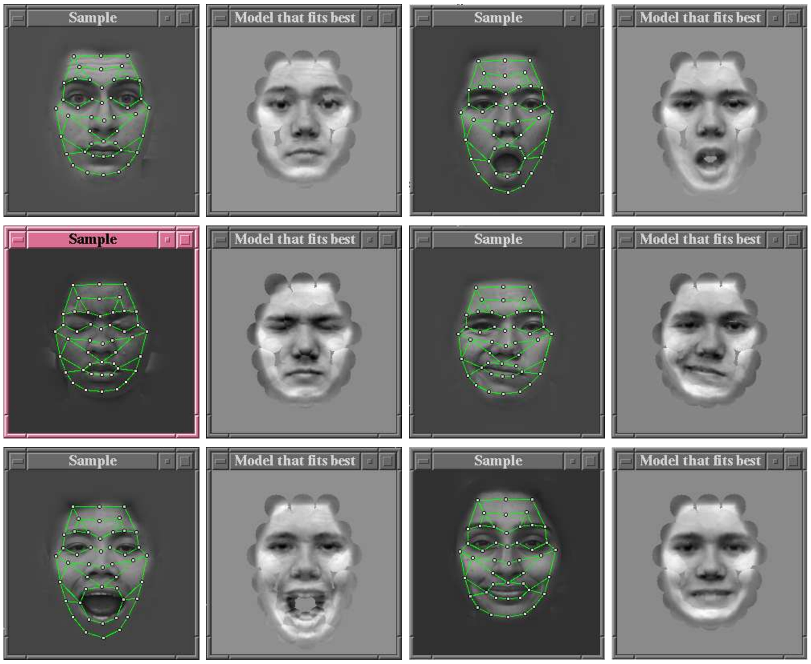


Fig. 7. FOM Matching for six different facial expressions. The sample images (first and third column) are shown with suppressed background and superimposed final graph position, while the correspondent image to the right is a reconstruction from the 4-person flexible object model.

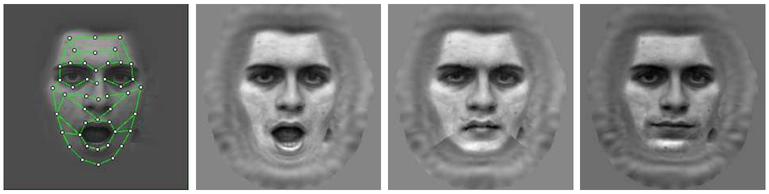


Fig. 8. Estimation of neutral expression using the FOM. From left to right are shown the original image with the best-matching graph, the image reconstructed from that graph, the estimated neutral expression using a person-indpendent FOM, and finally a reconstruction of the neutral-expression gallery image from its graph representation.

6 Conclusions

We have presented an extension of the established concept of Elastic Graph Matching. Instead of synthetically constructing a model for shape variation we empirically learn it from sample image sequences requiring only minimal assistance. The model describes flexible objects in terms of deformation in shape and in texture as well as a linear mapping between the two. Applications to facial gestures are investigated in exploratory experiments. As the model is based on the data format of EGM it is immediately applicable to image matching operations, as demonstrated. More extensive experiments like recognition tasks using a larger database and further applications are in progress.

Acknowledgements: Funding by the European Commission in the Research and Training Network MUHCI (HPRN-CT-2000-00111) and by the Deutsche Forschungsgemeinschaft (WU 314/2-1) is gratefully acknowledged.

References

1. Wiskott, L., Fellous, J.M., Krüger, N., von der Malsburg, C.: Face recognition by elastic bunch graph matching. IEEE Trans. Pattern Anal. Mach. Intell. **19** (1997) 775–779
2. Okada, K.: Analysis, Synthesis and Recognition of Human Faces with Pose Variations. PhD thesis, University of Southern California (2001)
3. Hong, H.: Analysis, Recognition and Synthesis of Facial Gestures. PhD thesis, University of Southern California (2000)
4. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision **60** (2004) 135 – 164
5. Maurer, T., von der Malsburg, C.: Tracking and learning graphs and pose on image sequences of faces. In: Proceedings of the 2nd INternational Conference on Automatic Face and Gesture Recognition (FG '96), IEEE Computer Society (1996) 76
6. Bartlett, M.S.: Face Image Analysis by Unsupervised Learning. Kluwer Academic Publishers (2001)
7. Fritzke, B.: A growing neural gas network learnes topologies. In: Advances in Neural Information Processing Systems 7. MIT Press (1995) 625–632
8. Kegl, B., Krzyzak, A., Linder, T., Zeger, K.: Learning and design of principal curves. IEEE Trans. Pattern Anal. Mach. Intell. **22** (2000) 281–297
9. Campbell, S.L., C.D. Meyer, J.: Generalized Inverses of Linear Transformations. Dover Publications (1991)
10. Pötzsch, M., Maurer, T., Wiskott, L., von der Malsburg, C.: Reconstruction from graphs labeled with responses of gabor filters. In: Proceedings of the ICANN 1996. (1996) 845–850