Recognizing Facial Expressions in Videos using a Facial Action Analysis-Synthesis Scheme*

Fadi Dornaika and Bogdan Raducanu
Computer Vision Center
08193 Bellaterra, Barcelona, SPAIN
{dornaika, bogdan}@cvc.uab.es

Abstract

In this paper, we propose a novel approach for facial expression analysis and recognition. The proposed approach relies on tracked facial actions provided by an appearance-based 3D face tracker. For each universal expression, a dynamical model for facial actions given by an auto-regressive process is learned from training data. We classify a given image in an unseen video into one of the universal facial expression categories using an analysis-synthesis scheme. This scheme uses all models and select the one that provides the most consistent synthesized spatio-temporal facial actions. The dynamical models can be utilized in the tasks of synthesis and prediction. Experiments using unseen videos demonstrated the effectiveness of the developed method.

1 Introduction

Facial expression plays an important role in cognition of human emotions. Basic facial expressions typically recognized by psychologists are happiness, sadness, fear, anger, disgust and surprise [9]. In the past, facial expression analysis was essentially a research topic for psychologists. However, recent progresses in image processing and pattern recognition have motivated significantly research works on automatic facial expression recognition [11].

The recognition of facial expressions in image sequences with significant head motion is a challenging problem. It is required by many applications such as human-computer interaction and computer graphics animation. To classify expressions in still images many techniques have been proposed such as Neural Nets [16], Gabor wavelets [3], and active appearance models [1]. The still images usually capture the apex of the expression, i.e. the instant at which the indicators of emotion are most marked. Recently, more attention has been given to modeling facial deformation in dynamic scenarios. Still image classifiers use feature vectors related to a single frame to perform classification. Temporal classifiers try to capture the temporal pattern in the sequence of feature vectors related to each frame such as the Hidden Markov Models based methods [7, 5, 15] and Dynamic Bayesian Networks [17]. In [5], parametric 2D flow models associated with the whole face as well as with the mouth, eyebrows, and eyes are first estimated. Then, mid-level predicates are inferred from these parameters. Finally, universal facial expressions are detected and recognized using the estimated predicates.

In this paper, we propose a method for recognizing emotions through facial expressions displayed in video sequences. We propose a novel scheme for facial expression recognition that is based on an appearance-based 3D face tracker. Our developed approach enables facial expression recognition using an analysis-synthesis scheme based on auto-regressive models. Although auto-regressive models have been widely used in the tasks of 2D tracking and synthesis, to the best of our knowledge they have not been used for facial expression recognition. The proposed approach proceeds as follows. First, a tracker provides the time-varying facial actions related to the lips and the eyebrows. Second, using learned auto-regressive models (each universal expression has a model) the facial actions are then temporally synthesized. Then similarity measures between the synthesized trajectories and the actual ones will decide the expression.

Compared to existing temporal facial expression methods our proposed approach has several advantages. First, unlike most expression recognition systems that require a frontal view of the face, our sys-

*This work was supported by the MEC project TIN2005-09026 and The Ramón y Cajal Program.

tem is view-independent since the used tracker simultaneously provides the 3D head pose and the facial actions. Second, it is texture-independent since the recognition scheme relies only on the estimated facial actions. Third, its learning phase is simple compared to other techniques (e.g., the Hidden Markov Models and Active Appearance Models), that is, we only need to fit second-order Auto-Regressive models to sequences of facial actions. As a result, even when the imaging conditions change the learned Auto-Regressive models need not to be recomputed. The rest of the paper is organized as follows. Section 2 summarizes our developed appearance-based 3D face tracker that we use to track the 3D head pose as well as the facial actions. Section 3 describes the proposed facial expression recognition. Section 4 provides some experimental results.

2 Simultaneous head and facial action tracking

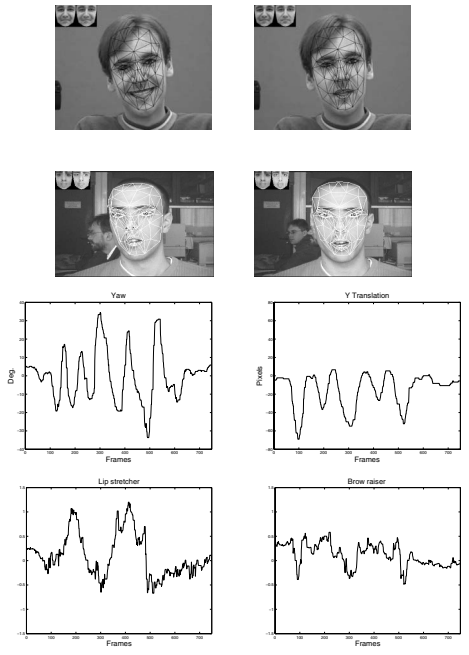


Figure 1. Top: Simultaneous head and facial action tracking results associated with two video sequences. **Bottom:** The yaw angle, the vertical translation, the lip stretcher, and the eye brow raiser associated with the second video sequence.

In our study, we use the 3D face model *Candide* [2]. This 3D deformable wireframe model is given by the 3D coordinates of the vertices $\mathbf{P}_i, i = 1, \ldots, n$ where n is the number of vertices. Thus, the shape up to a global scale can be fully described by the $3n$-vector $\mathbf{g}$ – the concatenation of the 3D coordinates of all vertices $\mathbf{P}_i$. The vector $\mathbf{g}$ can be written as:

$$\mathbf{g} = \mathbf{g}_s + \mathbf{A}\,\boldsymbol{\tau}_\mathbf{a} \tag{1}$$

where $\mathbf{g}_s$ is the static shape of the model, $\boldsymbol{\tau}_\mathbf{a}$ the facial action vector, and the columns of $\mathbf{A}$ are the Animation Units. In this study, we use six modes for the facial Animation Units (AUs) matrix $\mathbf{A}$, that is, the dimension of $\boldsymbol{\tau}_\mathbf{a}$ is 6. These modes are all included in the *Candide* model package. We have chosen the six following AUs: lower lip depressor, lip stretcher, lip corner depressor, upper lip raiser, eyebrow lowerer and outer eyebrow raiser. These AUs are enough to cover most common facial animations (mouth and eyebrow movements). Moreover, they are essential for conveying emotions. Many studies have shown that image regions associated with the mouth and the eyebrows are the most informative regions about the facial expression (e.g., [4, 5]).

Thus, the state of the 3D model is given by the 3D head pose (three rotations and three translations) and the vector $\boldsymbol{\tau}_\mathbf{a}$. This is given by the 12-vector $\mathbf{b}$:

$$\mathbf{b} = [\theta_x, \quad \theta_y, \quad \theta_z, \quad t_x, \quad t_y, \quad t_z, \quad \boldsymbol{\tau}_\mathbf{a}{}^T \,]^T \tag{2}$$

A cornerstone problem in facial expression recognition is the ability to track the local facial actions/deformations. In our work, we track the head and facial actions using our tracker [8]. This appearance-based tracker simultaneously computes the 3D head pose and the facial actions encapsulated in the vector $\mathbf{b}$ by minimizing a distance between the incoming warped frame and the current appearance of the face. This minimization is carried out using a Gauss-Newton method. The statistics of the appearance as well as the gradient matrix are updated every frame. This scheme leads to a fast and robust tracking algorithm. Figure 1 displays the tracking results associated with two video sequences. The first video [1] consists of 1000 frames, and depicts a subject engaged in conversation with another person. The second video consists of 750 frames, and depicts a subject featuring quite large head pose variations as well as large facial actions. The bottom of this figure displays the estimated value of the yaw angle, the vertical translation, the lip stretcher, and the brow raiser associated with the second sequence. Since the facial actions $\boldsymbol{\tau}_\mathbf{a}$ are highly correlated to the facial expressions, their time series representation can be utilized for inferring the facial expression in videos. This will be explained in the sequel. We stress the fact that since these actions are independent from the 3D

[1] www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/

head pose, our proposed facial expression recognition method will be view-independent.

3 Approach

In this section, we introduce a recognition scheme based on the dynamics of facial expressions. We argue that the facial expression can be inferred from the temporal representation of the tracked facial actions. For this purpose, we use continuous dynamical systems described by the facial action parameters $\boldsymbol{\tau}_{\mathbf{a}(t)}$.

Corresponding to each universal expression there is a dynamical model, supposed to be a Markov model of order K. It is a Gaussian Auto-Regressive Process (ARP) defined by

$$\boldsymbol{\tau}_{\mathbf{a}(t)} = \sum_{k=1}^{K} \mathbf{A}_k \, \boldsymbol{\tau}_{\mathbf{a}(t-k)} + \mathbf{d} + \mathbf{B} \, \mathbf{w}_{(t)} \qquad (3)$$

in which $\mathbf{w}_{(t)}$ is a vector of 6 (6 is the dimension of $\boldsymbol{\tau}_{\mathbf{a}(t)}$) independent random $\mathcal{N}(0,1)$ variables. The dynamical parameters of the model are: (i) deterministic parameters $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K$ and $\mathbf{d}$, and stochastic parameters $\mathbf{B}$, which determine the coupling of $\mathbf{w}_{(t)}$ into the vector $\boldsymbol{\tau}_{\mathbf{a}(t)}$. For convenience of notation, let $\mathbf{A}_{6\times 6K} = (\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K)$. It is worthwhile noting that the above model can be used for predicting the process from the previous K values. The predicted value at time t obeys a multivariate Gaussian centered at the deterministic value of (3) with $\mathbf{B}\mathbf{B}^T$ being its covariance matrix [6, 13, 14].

3.1 Learning

Given a training sequence $\boldsymbol{\tau}_{\mathbf{a}(1)}, \cdots, \boldsymbol{\tau}_{\mathbf{a}(T)}$, with $T > K$, belonging to the same universal expression, it is well-known that a Maximum Likelihood Estimator provides a closed-form solution for the parameters $\mathbf{A}, \mathbf{d}$, and $\mathbf{B}$.

In the sequel, we are interested in second order models, i.e. $K = 2$. The reason is twofold. First, these models are easy to estimate. Second, they are able to model complex motions. For example, these models have been used in [6] for learning the 2D motion dynamics of talking lips, beating hearts, and writing fingers.

For a second-order model, the model parameters reduce to two 6×6 matrices $\mathbf{A}_1, \mathbf{A}_2$, a 6-vector $\mathbf{d}$, and a 6×6 covariance matrix $\mathbf{C}$. Therefore, equation (3) reduces to:

$$\boldsymbol{\tau}_{\mathbf{a}(t)} = \mathbf{A}_1 \, \boldsymbol{\tau}_{\mathbf{a}(t-1)} + \mathbf{A}_2 \, \boldsymbol{\tau}_{\mathbf{a}(t-2)} + \mathbf{d} + \mathbf{B} \, \mathbf{w}_{(t)} \qquad (4)$$

We have built a second-order auto-regressive model for each universal expression: Surprise, Sadness, Joy, Disgust, and Anger. We have used two different training sets. The first training set was provided by the CMU data [12] which consist of 35 short videos depicting these five universal expressions. Each expression was performed by 7 persons (see Figure 2).

Notice that in the case of the CMU set, the auto-regressive models were computed by concatenating the facial actions associated with the 7 persons.

The second data set consisted of five 30-second videos. Each video sequence contains several cycles depicting a given universal expression. All these videos were performed by the same subject. Figure 3 shows the tracked facial actions associated with the three training 30-second videos.

It should be noticed that neutral expressions can present slight deformations even when the face of the subject seems expressionless. However, we have not used auto-regressive models for modelling the dynamics of these slight deformations. Instead we use the $\mathcal{L}1$ norm of the vector $\boldsymbol{\tau}_{\mathbf{a}(t)}$ to decide if the corresponding current frame depicts a neutral expression or not.

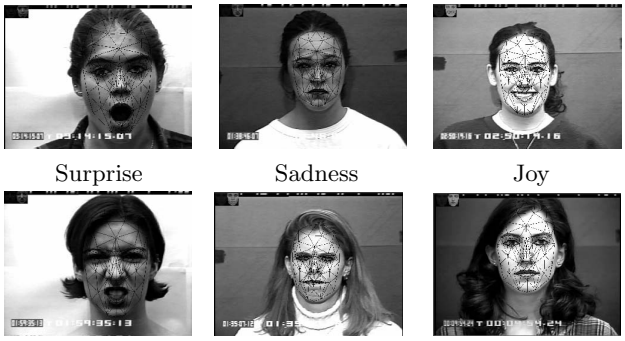


Figure 2. Six training videos from the CMU database. The first five images depict the high magnitude of the five basic expressions together with the fitted 3D deformable model.

3.2 Recognition

We infer the facial expression in videos by considering the vectors $\boldsymbol{\tau}_{\mathbf{a}(t)}$ within a temporal window of size T centered at the current frame t. These vectors are provided by the 3D face tracker. The expression for frame t is recognized using the following analysis-synthesis scheme. This is a two-step approach. In the first step, we locally synthesize the facial actions, $\hat{\boldsymbol{\tau}}_{\mathbf{a}(i)}$ within the temporal window using all auto-regressive

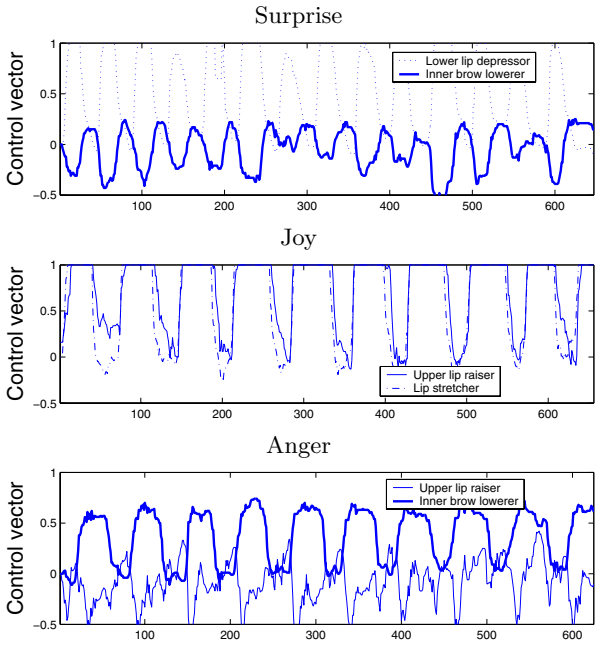


Figure 3. The tracked facial actions, $\boldsymbol{\tau}_{\mathbf{a}(t)}$, associated with three training videos. For a given plot, only two components are displayed.

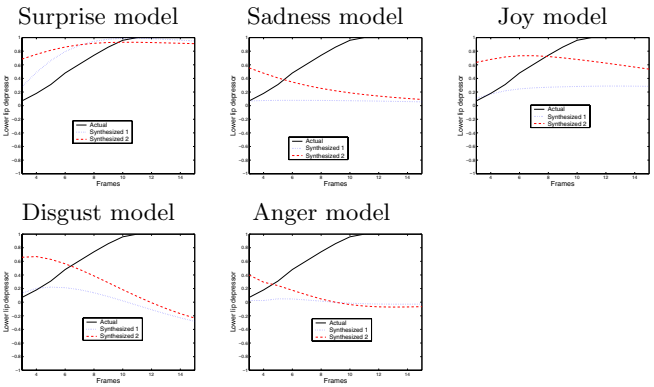


Figure 4. This example illustrates the analysis-synthesis scheme associated with the lower lip depressor parameter as a function of time (13 frames). Each graph corresponds to a given auto-regressive model. The actual parameter (solid curve) which corresponds to a tracked surprise transition is the same for all graphs. In this example, one can easily see how the synthesized trajectories are very similar to the actual one when the surprise auto-regressive model is used (top-left).

models and the actual tracked facial actions. In the second step, the model providing the most similar facial action trajectory to the actual one will decide the classification.

For the synthesis purpose, we utilize the deterministic part of (3). Recall that second-order auto-regressive models require two initial frames. Thus, in our synthesis process, we generate two synthesized facial action trajectories for each expression according to the choice of these initial values. The first trajectory is recursively generated from (3) where the two initial frames are set to their corresponding actual values. The second trajectory is recursively generated from (3) where the two initial frames are set to the local average value of the actual trajectory.

Let $\mathbf{r}$ be the actual facial action trajectory - the concatenation of the tracked facial actions $\boldsymbol{\tau}_{\mathbf{a}(t)}$ within the temporal window of size T. The dimension of $\mathbf{r}$ is $6T$. Let $\mathbf{s}$ be a synthesized facial action trajectory - a $6T$-vector. Since we have five auto-regressive models and two synthesis schemes then we have 10 synthesized trajectories $\mathbf{s}_{kl}, k = 1, \ldots, 5, l = 1, 2$ (recall that we have two synthesized trajectories for each universal expression).

Comparing the actual trajectory $\mathbf{r}$ with a synthesized one $\mathbf{s}_{kl}$ can be carried out using the cosine of the angle between the two vectors:

$$d_{kl} = \frac{\mathbf{r}^T \mathbf{s}_{kl}}{|\mathbf{r}|\,|\mathbf{s}_{kl}|} \tag{5}$$

Let $d_k = max(d_{kl}), l = 1, 2$. In other words, we only retain the synthesized trajectory that is the most consistent with the actual tracked one. Therefore, the most probable universal expression depicted in the current frame will be given by:

$$k = arg \max_k (d_k), \quad k = 1, \ldots, 5$$

The above similarity measures can be normalized. For example, the following normalization can be used:

$$p_k = \frac{e^{d_k}}{\sum_{j=1}^{5} e^{d_j}}$$

Figure 4 illustrates the analysis-synthesis scheme associated with the lower lip depressor parameter as a function of time (13 frames). Each graph corresponds to a given auto-regressive model. The solid curve corresponds to the actual facial parameter which corresponds to a tracked surprise transition. The solid curve is the same for all graphs. The dashed and dotted curves correspond to the synthesized parameter using the learned auto-regressive models: the dashed ones

correspond to the case where the initial conditions are set to the local average while the dotted ones to the case where the initial conditions are set to two actual values.

In this real example, one can easily see how the surprise auto-regressive-based synthesized trajectories are very similar to the actual trajectory (top-left).

4 Experimental results

Our experiments were performed on three video sequences. Each test video was acquired by a different camera and depicted a series of facial expressions performed by an unseen subject. In other words, the subject in each test video was different from those used for learning the auto-regressive models. In the first experiment, we have used a 748-frame-long test sequence. Eight frames of this sequence are shown in Figure 5. The bottom of this figure shows the normalized similarities associated with each universal expression obtained with the sequence using a sliding temporal window of 15 frames. The used auto-regressive models were built with the CMU data. By inspecting the original video we have found that all displayed expressions were correctly classified by the developed approach (Section 3) except for the disgust expression for which the approach provides a mixture of three expressions (see the similarity curves at frames 200 and 500). Note that the temporal window size should be greater than or equal to the minimum time needed by a an expression to go from the neutral configuration to a perceived expression.

In the second experiment, we used a 300-frame-long video sequence. For this sequence, we asked a subject to display several expressions arbitrarily (see Figure 6). The bottom of this figure shows the normalized similarities associated with each universal expression. In this case, the auto-regressive models were built with the second training data set (the five 30-second videos)). As can be seen, the algorithm has correctly detected the presence of the surprise, joy, and sadness expressions. Note that the mixture of expressions at transition is normal since the recognition is performed in a frame-wise manner.

In the third experiment, we have used a 325-frame-long video sequence. Figure 7 shows the recognition results associated with this video.

In order to quantify the recognition rate, we have used the 35 CMU videos for testing using the auto-regressive built with the second training data set. Table 1 shows the confusion matrix associated with the 35 test videos illustrating 7 persons. As can be seen, although the recognition rate was good (80%), it is not equal to 100%. This can be explained by the fact that the expression dynamics could be highly person-dependent. Recall that the used auto-regressive models are built using data associated with one single person. Notice that the human ceiling in correctly classifying facial expressions into the six basic emotions has been established at 91.7% by Ekman & Friesen [10].

Figure 8 summarizes the joy test data (CMU data) used for the confusion matrix computation. This figure displays the value of the cosine as defined by (5) for 7 test videos concatenated into one single sequence.

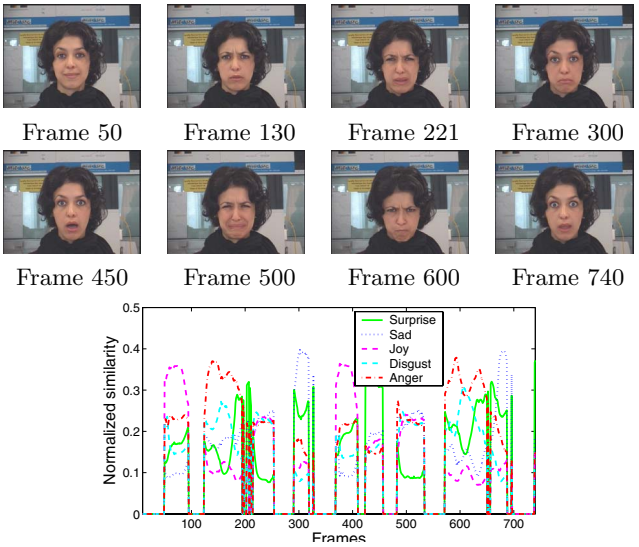


Figure 5. Top: Eight frames associated with a 748-frame-long test sequence. **Bottom:** The similarity measure computed for each universal expression and for each non-neutral frame of the sequence.

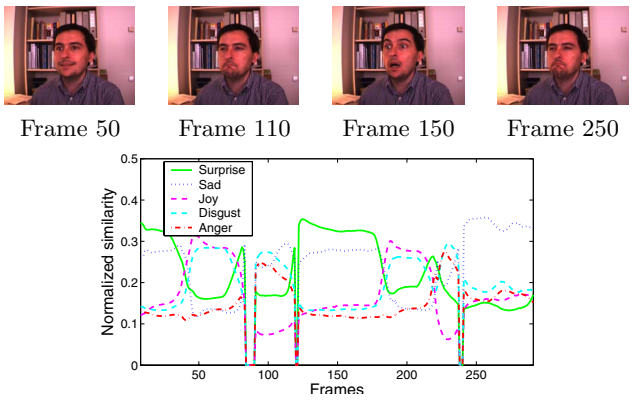
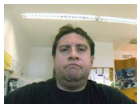
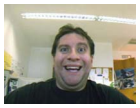


Figure 6. Top: Four frames (50, 110, 150, and 250) associated with a 300-frame-long test sequence. **Bottom:** The similarity measure computed for each universal expression and for each non-neutral frame of the sequence.


Frame 75 (Anger)


Frame 110 (Joy)

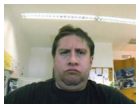

Frame 250 (Sadness)

Figure 7. Three frames associated to the third test sequence. The recognition results are indicated in parentheses.

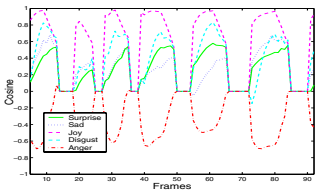


Figure 8. The cosine angle associated with 7 short video illustrating joy expressions performed by seven different persons (the videos are concatenated into one single video). As can be seen, the maximum of the cosine correctly indicates the joy expression.

	Surp.	Sad.	Joy	Disg.	Ang.
Surp.	7	0	0	0	0
Sad.	0	7	0	5	0
Joy	0	0	7	0	0
Disg.	0	0	0	2	2
Ang.	0	0	0	0	5

Table 1. Confusion matrix for the facial expression classifier associated with 35 test videos (CMU data).

5 Conclusion

This paper described a view- and texture-independent approach to facial expression analysis and recognition. The approach relies on a analysis-synthesis scheme by which the spatio-temporal facial actions are locally synthesized by means of learned auto-regressive models and then compared to the actual parameters. Based on this comparison a frame can be classified into one universal expression. We found out that the recognition performance depends largely on the representational capacity of the training set. Our performed experiments confirm the general trend of dynamical classifiers that seem to be very accurate for a given person but the accuracy decreases when one attempts to generalize them. Future research will explore advanced methods allowing to take into account the inter-person dynamics within the framework of auto-regressive models.

References

[1] B. Abboud, F. Davoine, and M. Dang. Facial expression recognition and synthesis based on an appearance model. *Signal Processing: Image Communication*, 19(8):723–740, 2004.

[2] J. Ahlberg. CANDIDE-3 - an updated parametrized face. Technical Report LiTH-ISY-R-2326, Department of Electrical Engineering, Linköping University, Sweden, 2001.

[3] M. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan. Machine learning methods for fully automatic recognition of facial expressions and facial actions. In *IEEE Int. Conference on Systems, Man and Cybernetics*, 2004.

[4] J. N. Bassili. Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37:2049–2059, 1979.

[5] M. J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1):23–48, 1997.

[6] A. Blake and M. Isard. *Active Contours*. Springer-Verlag, 2000.

[7] I. Cohen, N. Sebe, A. Garg, L. Chen, and T. Huang. Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding.*, 91(1-2):160–187, 2003.

[8] F. Dornaika and F. Davoine. On appearance based face and facial action tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, In press.

[9] P. Ekman. Facial expressions of emotion: an old controversy and new findings. *Philosophical Transactions of the Royal Society, London*, B335:63–69, 1992.

[10] P. Ekman and W. Friesen. *Pictures of Facial Affect*. Consulting Psychologists Press, Palo Alto, CA, USA, 1976.

[11] B. Fasel and J. Luettin. Automatic facial expression analysis: A survey. *Pattern Recognition*, 36(1):259–275, 2003.

[12] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *International Conference on Automatic Face and Gesture Recognition*, pages 46–53, Grenoble, France, March 2000.

[13] L. Ljung. *System Identification: Theory for the User*. Prentice Hall, 1987.

[14] B. North, A. Blake, M. Isard, and J. Rittscher. Learning and classification of complex dynamics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):1016–1034, 2000.

[15] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[16] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:97–115, 2001.

[17] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):699–714, 2005.