Evaluation of Expression Recognition Techniques

Ira Cohen[1], Nicu Sebe[2,3], Yafei Sun[3], Michael S. Lew[3], Thomas S. Huang[1]

[1]Beckman Institute, University of Illinois at Urbana-Champaign, USA
[2]Faculty of Science, University of Amsterdam, The Netherlands
[3]Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands

Abstract. The most expressive way humans display emotions is through facial expressions. In this work we report on several advances we have made in building a system for classification of facial expressions from continuous video input. We introduce and test different Bayesian network classifiers for classifying expressions from video. In particular we use Naive-Bayes classifiers and to learn the dependencies among different facial motion features we use Tree-Augmented Naive Bayes (TAN) classifiers. We also investigate a neural network approach. Further, we propose an architecture of hidden Markov models (HMMs) for automatically segmenting and recognizing human facial expression from video sequences. We explore both person-dependent and person-independent recognition of expressions and compare the different methods.

1 Introduction

It is argued that to truly achieve effective human-computer intelligent interaction (HCII), there is a need for the computer to be able to interact naturally with the user, similar to the way human-human interaction takes place. Humans interact with each other mainly through speech, but also through body gestures, to emphasize a certain part of the speech and display of emotions. One of the important way humans display emotions is through facial expressions.

Ekman and Friesen [1] developed the Facial Action Coding System to code facial expressions where movements on the face are described by a set of action units (AUs). Ekman's work inspired many researchers to analyze facial expressions by means of image and video processing. By tracking facial features and measuring the amount of facial movement, they attempt to categorize different facial expressions. Recent work on facial expression analysis and recognition [2–8] has used these "basic expressions" or a subset of them. In [9], Pantic and Rothkrantz provide an in depth review of many of the research done in automatic facial expression recognition in recent years. These methods are similar in that they first extract features, then these features are used as inputs into a classification system, and the outcome is one of the preselected emotion categories. They differ mainly in the features extracted from the videos and in the classifiers used to distinguish between the different emotions.

Our work focuses on the design of the classifiers used for performing the recognition following extraction of features using our real time face tracking system. We describe classification schemes in two types of settings: dynamic and 'static' classification.

The 'static' classifiers use feature vectors related to a single frame to perform classification (e.g., Neural networks, Bayesian networks). More specifically, we use

two types of Bayesian network classifiers: Naive Bayes, in which the features are assumed independent given the class, and the Tree-Augmented Naive Bayes classifier (TAN). While Naive-Bayes classifiers are often successful in practice, they use the very strict and often unrealistic independence assumption. To account for this, we use the TAN classifiers which have the advantage of modeling dependencies between the features without much added complexity compared to the Naive-Bayes classifiers. We were also interested in using a neural network approach. Dynamic classifiers take into account the temporal pattern in displaying facial expression. We propose a multi-level HMM classifier, combining the temporal information which allows not only to perform the classification of a video segment to the corresponding facial expression, as in the previous works on HMM based classifiers [6, 8], but also to automatically segment an arbitrary long video sequence to the different expressions segments without resorting to heuristic methods of segmentation.

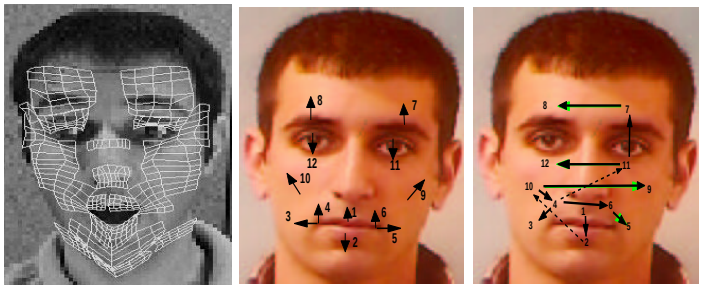
An important aspect is that while the 'static' classifiers are easier to train and implement, the dynamic classifiers require more training samples and many more parameters to learn.

2 Face Tracking and Feature Extraction

The face tracking we use in our system is based on a system developed by Tao and Huang [10] called the Piecewise Bézier Volume Deformation tracker. This face tracker uses a model-based approach where an explicit 3D wireframe model of the face is constructed. In the first frame of the image sequence, landmark facial features such as the eye corners and mouth corners are selected interactively. The generic face model is then warped to fit the selected facial features. The face model consists of 16 surface patches embedded in Bézier volumes. The surface patches defined in this way are guaranteed to be continuous and smooth. The shape of the mesh can be changed by changing the locations of the control points in the Bézier volume.

Once the model is constructed and fitted, head motion and local deformations of the facial features such as the eyebrows, eyelids, and mouth can be tracked. First the 2D image motions are measured using template matching between frames at different resolutions. Image templates from the previous frame and from the very first frame are both used for more robust tracking. The measured 2D image motions are modeled as projections of the true 3D motions onto the image plane. From the 2D motions of many points on the mesh, the 3D motion can be estimated by solving an overdetermined system of equations of the projective motions in the least squared sense. Figure 1(a) shows an example from one frame of the wireframe model overlayed on a face being tracked.

The recovered motions are represented in terms of magnitudes of some predefined motion of various facial features. Each feature motion corresponds to a simple deformation on the face, defined in terms of the Bézier volume control parameters. We refer to these motions vectors as Motion-Units (MU's). Note that they are similar but not equivalent to Ekman's AU's and are numeric in nature, representing not only the activation of a facial region, but also the direction and intensity of the motion. The 12 MU's used in the face tracker are shown in Figure 1(b). They are used as the basic features for the classification scheme described in the next sections.



(a) (b) (c)

Fig. 1. (a) The wireframe model, (b) The facial motion measurements, (c) The learned TAN structure for the facial features. Dashed lines represent links that are relatively weaker than the others.

3 The Static Approach

We use Bayesian network classifiers and a neural network for recognizing facial expressions given the tracking results provided by the face tracking algorithm. Our classifiers are 'static' in the sense that their features are tracking results at each point in time.

3.1 Bayesian Networks for Expression Recognition

A Naive-Bayes classifier is a probabilistic classifier in which the features are assumed independent given the class. Although the Naive-Bayes model does not reflect in many cases the true underlying model generating the data, it is still observed to be successful as a classifier in practice. The reason for the Naive-Bayes model's success as a classifier is attributed to the small number of parameters needed to be estimated, thus offsetting the large modeling bias with a small estimation variance [11].

Given a Bayesian network classifier with parameter set Θ, the optimal classification rule under the maximum likelihood (ML) framework to classify an observed feature vector of n dimensions, $X \in R^n$, to one of $|C|$ class labels, $c \in \{1, ..., |C|\}$, is given as:

$$\hat{c} = \underset{c}{argmax} \;\; P(X|c;\Theta) \tag{1}$$

Based on the observation that the strong independence assumption may seem unreasonable for our application (the facial motion measurements are highly correlated when humans display emotions), we decided to go beyond the Naive-Bayes assumption. Therefore, we wanted to find a way to search for a structure that captures the dependencies among the features. Of course, to attempt to find all the dependencies is an NP-complete problem. So, we restricted ourselves to a smaller class of structures called the Tree-Augmented-Naive Bayes (TAN) classifiers [12]. The joint probability distribution is factored to a collection of conditional probability distributions of each node in the graph.

In the TAN classifier structure the class node has no parents and each feature has as parents the class node and at most one other feature, such that the result is a tree structure for the features. Friedman et al. [12] proposed using the TAN model as a classifier, to enhance the performance over the simple Naive-Bayes classifier. The existence of an efficient algorithm to compute the best TAN model makes it a good candidate in the search for a better structure over the simple NB. This

method is using the modified Chow-Liu algorithm [13] for constructing tree augmented Bayesian networks [12]. The algorithm finds the tree structure among the features that maximizes the likelihood of the data by computation of the pairwise class conditional mutual information among the features and building a maximum weighted spanning tree (MWST) using the pairwise mutual information as the weights of the arcs in the tree.

For facial expression recognition, the learned TAN structure can provide additional insight on the interaction between facial features in determining facial expressions. Figure 1(c) shows a learned tree structure of the features (our Motion Units) learned using our database of subjects displaying different facial expressions (more details on the experiments are in Section 5). The arrows are from parents to children MUs. From the tree structure we see that the TAN learning algorithm produced a structure in which the bottom half of the face is almost disjoint from the top portion, except for a weak link between MU 4 and MU 11.

We have recently showed that using NB or TAN classifiers can achieve good results for facial expression recognition, where the choice between each structure depends mainly on the size of the training set [14].

3.2 Neural Network Approach

In a neural network based classification approach, a facial expression is classified according to the categorization process the network learned during the training phase. In our implementation, we used the approach proposed by Padgett and Cottrell [15]. For classification of the tracking results provided by the face tracking algorithm into one of 6 basic categories (happiness, sadness, anger, fear, surprise, and disgust) plus Neutral emotion category, we use a back-propagation neural network. The input to the network consists of the 12 MU extracted by the face tracking algorithm. The hidden layer of the NN contains 10 nodes and employs a nonlinear Sigmoid activation function [15]. The output layer of the NN contains 7 units, each of which corresponds to one emotion category.

For training and the testing of the neural network we use the same training and testing data as in the case of Bayesian Networks. See Section 5 for more details.

4 The Dynamic Approach

The dynamic approach employes classifiers that can use temporal information to discriminate between different expressions. The logic behind using the temporal information is that expressions have a unique temporal pattern. When recognizing expressions from video, the use of temporal information can lead to more robust and accurate classification results compared to methods that are 'static'.

4.1 Expression Recognition Using Multi-level HMM

The method we propose automatically segments the video to the different facial expression sequences, using a multi-level HMM structure. To solve the segmentation problem and enhance the discrimination between the classes we propose the achitecture shown in Figure 2. This architecture performes automatic segmentation and recognition of the displayed expression at each time instance. The motion features are continuously used as input to the six emotion-specific HMMs. The state sequence of each of the HMMs is decoded and used as the observation vector for the high-level Markov model. This model consists of seven states, one for each of

the six emotions and one for Neutral. The Neutral state is necessary as for the large portion of time, there is no display of emotion on a person's face. In this implementation of the system, the transitions between emotions are imposed to pass through the Neutral state since our training data consists of facial expression sequences that always go through the Neutral state.

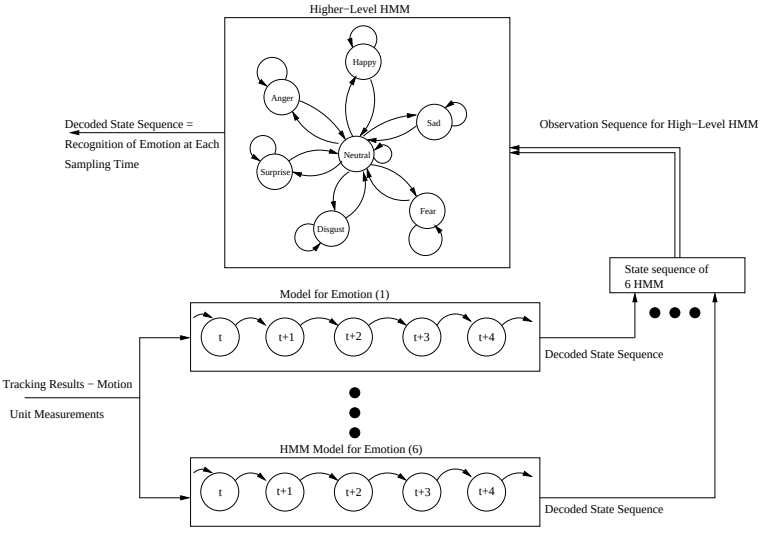


Fig. 2. Multi-level HMM architecture for automatic segmentation and recognition of emotion.

The recognition of the expression is done by decoding the state that the high-level Markov model is in at each point in time since the state represents the displayed emotion. The training procedure of the system is as follows:

- Train the emotion-specific HMMs using a hand segmented sequence.
- Feed all six HMMs with the continuous (labeled) facial expression sequence. Each expression sequence contains several instances of each facial expression with *neutral* instances separating the emotions.
- Obtain the state sequence of each HMM to form the six-dimensional observation vector of the higher-level Markov model, i.e., $O_t^h = [q_t^{(1)},...,q_t^{(6)}]^T$, where $q_t^{(i)}$ is the state of the i^{th} emotion-specific HMM. The decoding of the state sequence is done using the Viterbi algorithm [16].
- Learn the probability observation matrix for each state of the high-level Markov model using $P(q_j^{(i)}|S_k) = \{$expected frequency of model i being in state j given that the true state was $k\}$, and

$$B^{(h)} = \{b_k(O_t^h)\} = \left\{\prod_{i=1}^{6}(P(q_j^{(i)}|S_k)\right\} \tag{2}$$

  where $j \in (1,$*Number of States for Lower Level HMM).*
- Compute the transition probability $A = \{a_{kl}\}$ of the high-level HMM using the frequency of transiting from each of the six emotion classes to the *neutral*

state in the training sequences and from the *neutral* state to the other emotion states. For notation, the *neutral* state is numbered 7 and the other states are numbered as in the previous section. All the transition probabilities could also be set using expert knowledge.
- Set the initial probability of the high-level Markov model to be 1 for the *neutral* state and 0 for all other states. This forces the model to always start at the *neutral* state and assumes that a person will display a *neutral* expression in the beginning of any video sequence. This assumption is made just for simplicity of the testing.

The steps followed during the testing phase are very similar to the ones followed during training. The face tracking sequence is used as input into the lower-level HMMs and a decoded state sequence is obtained using the Viterbi algorithm. The decoded lower-level state sequence O_t^h is used as input to the higher-level HMM and the observation probabilities are computed using Eq. (2). Note that in this way of computing the probability, it is assumed that the state sequences of the lower-level HMMs are independent given the true labeling of the sequence. This assumption is reasonable since the HMMs are trained independently and on different training sequences. In addition, without this assumption, the size of B will be enormous, since it will have to account for all possible combinations of states of the six lower-level HMMs, and it would require a huge amount of training data.

Using the Viterbi algorithm again for the high-level Markov model, a most likely state sequence is produced. The state that the HMM was in at time t corresponds to the expressed emotion in the video sequence at time t. To make the classification result robust to undesired fast changes, a smoothing of the state sequence is done by preserving the actual classification result if the HMM did not stay in a particular state for more than T times, where T can vary between 1 and 15 samples (assuming a 30Hz sampling rate). The introduction of the smoothing factor T will cause a delay in the decision of the system, but of no more than T sample times.

5 Experiments

We use two different databases, a database collected by us and the Cohn-Kanade database [17]. The first is a database of subjects that were instructed to display facial expressions corresponding to the six types of emotions. All the tests of the algorithms are performed on a set of five people, each one displaying six sequences of each one of the six emotions, starting and ending at the Neutral expression. Each video sequence was used as the input to the face tracking algorithm. The sampling rate was 30Hz, and a typical emotion sequence is about 70 samples long ($\sim$2s).

We use our database in two types of experiments. First we performed person dependent experiments, in which part of the data for each subject was used as training data, and another part as test data. Second, we performed person independent experiments, in which we used the data of all but one person as training data, and tested on the person that was left out.

The Cohn-Kanade database [17] consists of expression sequences of subjects, starting from a Neutral expression and ending in the peak of the facial expression. There are 104 subjects in the database. Because for some of the subjects, not all of the six facial expressions sequences were available to us, we used a subset of 53 subjects, for which at least four of the sequences were available. For each person there are on average 8 frames for each expression, which makes insufficient data to

perform person dependent tests. Also, the fact that each sequence ends in the peak of the facial expression makes the use of our dynamic multi-level HMM classifier impractical since in this case each sequence counts for an incomplete temporal pattern.

For the frame based methods, we measure the accuracy with respect to the classification result of each frame, where each frame in the video sequence was manually labeled to one of the expressions (including Neutral). The accuracy for the temporal based methods is measured with respect to the misclassification rate of an expression sequence, not with respect to each frame.

5.1 Results Using Our Database

A person-dependent test is first tried. Tables 1 shows the recognition rate of each subject and the average recognition rate of the classifiers.

Subject	NB	TAN	NN	HMM
1	81.69%	85.94%	82.37%	80.05%
2	84.54%	89.39%	85.23%	85.71%
3	83.05%	86.58%	81.17%	80.56%
4	79.25%	82.84%	80.05%	88.89%
5	71.74%	71.78%	75.23%	77.14%
Average	80.05%	83.31%	80.81%	82.46%

Table 1. Person-dependent facial expression recognition rates

The fact that subject 5 was poorly classified can be attributed to the inaccurate tracking result and lack of sufficient variability in displaying the emotions.

It is also important to observe that taking into account the dependencies in the features (the TAN model) gives significantly improved results. Also, the neural network approach gives comparable results to all the other methods.

The confusion matrix for the TAN classifier is presented in Table 2. The analysis of the confusion between different emotions shows that most of the confusion of the classes is with the Neutral class. This can be attributed to the arbitrary labeling of each frame in the expression sequence. The first and last few frames of each sequence are very close to the Neutral expression and thus are more prone to become confused with it. We also see that most expression do not confuse with Happy.

Emotion	Neutral	Happy	Anger	Disgust	Fear	Sad	Surprise
Neutral	79.58	1.21	3.88	2.71	3.68	5.61	3.29
Happy	1.06	87.55	0.71	3.99	2.21	1.71	2.74
Anger	5.18	0	85.92	4.14	3.27	1.17	0.30
Disgust	2.48	0.19	1.50	83.23	3.68	7.13	1.77
Fear	4.66	0	4.21	2.28	83.68	2.13	3.00
Sad	13.61	0.23	1.85	2.61	0.70	80.97	0
Surprise	5.17	0.80	0.52	2.45	7.73	1.08	82.22

Table 2. Person-dependent confusion matrix using the TAN classifier

The confusion matrices for the HMM based classifiers (described in details in [18]) show similar results, with Happy achieving near 100%, and Surprise approximately 90%.

We saw that a good recognition rate was achieved when the training sequences were taken from the same subject as the test sequences. A more challenging application is to create a system which is person-independent. For this test all of the sequences of one subject are used as the test sequences and the sequences of the remaining four subjects are used as training sequences. This test is repeated five times, each time leaving a different person out (leave-one-out cross-validation). Table 3 shows the recognition rate of the test for all classifiers. In this case, the recognition rates are lower compared with the person-dependent results. This means that the confusions between subjects are larger than those within the same subject.

Classifier	NB	TAN	NN	Multilevel HMM
Recognition rate	64.77%	66.53%	66.44%	58.63%

Table 3. Recognition rate for person-independent test.

The TAN classifier provides the best results. One of the reasons for the misclassifications is the fact that the subjects are very different from each other (three females, two males, and different ethnic backgrounds); hence, they display their emotion differently. Although it appears to contradict the universality of the facial expressions as studied by Ekman and Friesen [1], the results show that for practical automatic emotion recognition, consideration of gender and race play a role in the training of the system.

Table 4 shows the confusion matrix for the TAN classifier. We see that Happy, Fear, and Surprise are detected with high accuracy, and other expressions are greatly confused mostly with Neutral. Here the differences in the intensity of the expressions among the different subjects played a significant role in the confusion among the different expressions.

Emotion	Neutral	Happy	Anger	Disgust	Fear	Sad	Surprise
Neutral	76.95	0.46	3.39	3.78	7.35	6.53	1.50
Happy	3.21	77.34	2.77	9.94	0	2.75	3.97
Anger	14.33	0.89	62.98	10.60	1.51	9.51	0.14
Disgust	6.63	8.99	7.44	52.48	2.20	10.90	11.32
Fear	10.06	0	3.53	0.52	73.67	3.41	8.77
Sad	13.98	7.93	5.47	10.66	13.98	41.26	6.69
Surprise	4.97	6.83	0.32	7.41	3.95	5.38	71.11

Table 4. Person-independent average confusion matrix using the TAN classifier

5.2 Results Using the Cohn-Kanade Database

For this test we first divided our database in 5 sets which contain the sequences corresponding to 10 or 11 subjects (three sets with 11 subjects, two sets with 10 subjects). We used the sequences from a set as test sequences and the remaining sequences were used as training sequences. This test was repeated five times, each time leaving a different set out (leave-one-out cross-validation). Table 5 shows the recognition rate of the test for all classifiers. Note that the results obtained with this database are much better than the ones obtained with our database. This is because in this case we have more training data. For training we had available the data from more than 40 different persons. Therefore, the learned model is more accurate and can achieve better classification rates when using the test data.

Classifier	NB	TAN	NN
Recognition rate	68.14%	73.22%	73.81%

Table 5. Recognition rates for Cohn-Kanade database.

In average the best results were obtained using the NN followed by TAN and NB. The confusion matrix for the NN classifier is presented in Table 6. In this case, Surprise was detected with over 91% accuracy and Happy with over 77% accuracy. The other expressions are greatly confused with each other.

Emotion	Neutral	Happy	Anger	Disgust	Fear	Sad	Surprise
Neutral	77.27	0.21	2.89	9.09	2.27	5.37	2.89
Happy	0	77.87	10.66	0.82	10.66	0	0
Anger	2.04	1.36	74.83	20.41	0	1.36	0
Disgust	1.14	3.41	7.95	73.86	1.14	11.36	1.14
Fear	2.86	27.62	0.95	3.81	63.81	0.95	0
Sad	4.20	7.94	6.15	18.32	4.35	57.14	11.90
Surprise	0	0	0	0	1.01	7.07	91.92

Table 6. Person-independent average confusion matrix using the NN classifier

6 Summary and Discussion

In this work, we presented several methods for expression recognition from video. Our intention was to perform an extensive evaluation of our methods using static and dynamic classification.

In the case of 'static' classifiers the idea was to classify each frame of a video to one of the facial expressions categories based on the tracking results of that frame. The classification in this case was done using Bayesian networks classifiers and neural networks. A legitimate question is, "Is it always possible to learn the TAN structure from the data and use it in classification?" Provided that there is sufficient training data, the TAN structure indeed can be extracted and used in classification. However, when the data is insufficient the learned structure is unreliable and the use of the Naive-Bayes classifier is recommended. Also, it is important to observe that the neural network approach provided similar results compared to the Bayesian Networks approach.

In the case of dynamic classifiers the temporal information was used to discriminate different expressions. The idea is that expressions have a unique temporal pattern and recognizing these patterns can lead to improved classification results. This was done using the multi-level HMM architecture which does not rely on any pre-segmentation of the video stream.

When one should use a dynamic classifier versus a 'static' classifier? This is a difficult question to ask. It seems, both from intuition and from our results, that dynamic classifiers are more suited for systems that are person dependent due to their higher sensitivity not only to changes in appearance of expressions among different individuals, but also to the differences in temporal patterns. 'Static' classifiers are easier to train and implement, but when used on a continuous video sequence, they can be unreliable especially for frames that are not at the peak of an expression. Another important aspect is that the dynamic classifiers are more complex, therefore they require more training samples and many more parameters to learn compared with the static approach. A hybrid of classifiers using expression dynamics and static classification is the topic of our future research.

References

1. Ekman, P., Friesen, W.: Facial Action Coding System: Investigator's Guide. Consulting Psychologists Press, Palo Alto, CA (1978)
2. Ueki, N., Morishima, S., Yamada, H., Harashima, H.: Expression analysis/synthesis system based on emotion space constructed by multilayered neural network. Systems and Computers in Japan **25** (1994) 95–103
3. Lanitis, A., Taylor, C., Cootes, T.: A unified approach to coding and interpreting face images. In: Proc. 5th International Conference on Computer Vision (ICCV'95). (1995) 368–373
4. Rosenblum, M., Yacoob, Y., Davis, L.: Human expression recognition from motion using a radial basis function network architecture. IEEE Transactions on Neural Network **7** (1996) 1121–1138
5. Essa, I., Pentland, A.: Coding, analysis, interpretation, and recognition of facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 757–763
6. Otsuka, T., Ohya, J.: Recognizing multiple persons' facial expressions using HMM based on automatic extraction of significant frames from image sequences. In: Proc. Int. Conf. on Image Processing (ICIP'97). (1997) 546–549
7. Donato, G., Bartlett, M., Hager, J., Ekman, P., Sejnowski, T.: Classifying facial actions. IEEE Transactions on Pattern Analysis and Machine Intelligence **21** (1999) 974–989
8. Oliver, N., Pentland, A., Bérard, F.: LAFTER: A real-time face and lips tracker with facial expression recognition. Pattern Recognition **33** (2000) 1369–1382
9. Pantic, M., Rothkrantz, L.: Automatic analysis of facial expressions: The state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 1424–1445
10. Tao, H., Huang, T.: Connected vibrations: A modal analysis approach to non-rigid motion tracking. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'98). (1998) 735–740
11. Friedman, J.: On bias, variance, 0/1-loss, and the curse-of-dimensionality. Data Mining and Knowledge Discovery **1** (1997) 55–77
12. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine Learning **29** (1997) 131–163
13. Chow, C., Liu, C.: Approximating discrete probability distribution with dependence trees. IEEE Transactions on Information Theory **14** (1968) 462–467
14. Cohen, I., Sebe, N., Chen, L., Huang, T.: Facial expression recognition from video sequences: Temporal and static modeling. to appear in Computer Vision and Image Understanding (2003)
15. Padgett, C., Cottrell, G.: Representing face images for emotion classification. In: Conf. Advances in Neural Information Processing Systems. (1996) 894–900
16. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech processing. Proceedings of IEEE **77** (1989) 257–286
17. Kanade, T., Cohn, J., Tian, Y.: Comprehensive database for facial expression analysis (2000)
18. Cohen, I.: Automatic facial expression recognition from video sequences using temporal information. In: MS Thesis, University of Illinois at Urbana-Champaign, Dept. of Electrical Engineering (2000)