FACIAL EXPRESSION RECOGNITION FROM VIDEO SEQUENCES

Ira Cohen[1], *Nicu Sebe*[2], *Ashutosh Garg*[1], *Michael S. Lew*[2], *Thomas S. Huang*[1]

[1]University of Illinois at Urbana-Champaign, USA, {iracohen,ashutosh,huang}@ifp.uiuc.edu

[2]Leiden Institute of Advanced Computer Science, The Netherlands, {nicu, mlew}@liacs.nl

ABSTRACT

Recognizing human facial expression and emotion by computer is an interesting and challenging problem. In this paper we propose a method for recognizing emotions through facial expressions displayed in video sequences. We introduce a Tree-Augmented-Naive Bayes (TAN) classifier that learns the dependencies between the facial features and we provide an algorithm for finding the best TAN structure. Our person-dependent and person-independent experiments show that using this TAN structure provides significantly better results than using simpler NB-classifiers.

1. INTRODUCTION

In recent years there has been a growing interest in improving all aspects of the interaction between humans and computers. This emerging field has been a research interest for scientists from several different scholastic tracks, i.e., computer science, engineering, psychology, and neuroscience. These studies focus not only on improving computer interfaces, but also on improving the actions the computer takes based on feedback from the user. 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. Emotions are displayed by visual, vocal, and other physiological means. There is a growing amount of evidence showing that emotional skills are part of what is called "intelligence" [7].

Ekman and Friesen [5] developed the most comprehensive system for synthesizing facial expressions based on what they call Action Units (AU). In the early 1990s the engineering community started to use these results to construct automatic methods for recognizing emotions from facial expressions in images or video [9, 13, 10, 1]. Work on recognition of emotions from voice and video has been recently suggested and shown to work by Chen, et al. [2], and De Silva, et al. [4]. Sebe, et al. [11] proposed an emotion recognition method using a Naive Bayes (NB) model. They proposed a framework to choose the model distribution for each emotion (class) according to the available ground truth. Using this framework they showed that using a Cauchy model assumption provides better classification rate than using a Gaussian model assumption.

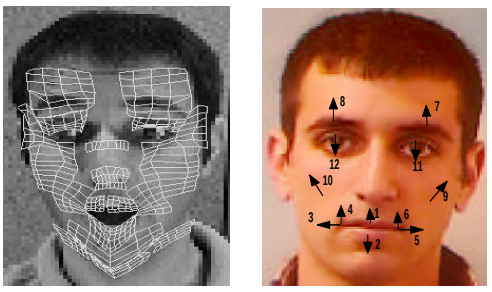
We propose a method for recognizing the emotions through facial expressions displayed in a video sequence. Our method is similar with the one proposed by Sebe, et al. [11] in the sense that we use the same features and we classify each frame of the video to a facial expression based on these features computed for that time frame. The novelty of this work is in going beyond the NB assumption. While in the NB assumption the features are considered to be independent, here we show that is beneficial to use the inherent dependencies that are present between the facial features.

The rest of the paper is organized as follows. Section 2 presents the features we use for the facial expression recognition. In Section 3 we introduce the Tree-Augmented-Naive Bayes (TAN) classifier which incorporates the dependencies between features and we present a method for learning the best TAN structure. In Section 4 we apply the theoretical results from Section 3 to determine the influence of the model assumption on the emotion classification results. Conclusions are given in Section 5.

2. FACE TRACKING AND FEATURE EXTRACTION

The very basis of any recognition system is extracting the best features to describe the physical phenomena. As such, categorization of the visual information revealed by facial expression is a fundamental step before any recognition of facial expressions can be achieved. First a model of the facial muscle motion corresponding to different expressions has to be found. This model has to be generic enough for most people if it is to be useful in any way.

The best known such model is given in the study by Ekman and Friesen [5], known as the Facial Action Coding System (FACS). Ekman has since argued that emotions are linked directly to the facial expressions and that there are six basic "universal facial expressions" corresponding to happiness, surprise, sadness, fear, anger, and disgust. The FACS codes the facial expressions as a combination of facial movements known as action units (AUs). The AUs have some relation to facial muscular motion and were defined based on anatomical knowledge and by studying videotapes of how the face changes its appearance. Ekman defined 46 such action units to correspond to each independent motion of the face. In our work, we consider a simplified model proposed by Tao and Huang [12] which uses an explicit 3D wireframe model of the face. The face model consists of 16 surface patches embedded in Bézier volumes. Figure 1 shows the wireframe model and the 12 facial motion measurements being measured for facial expression recognition, where the arrow represents the motion direction away from the neutral position of the face. The 12 features we use correspond to the magnitude of the 12 facial motion measurements defined in the face model and the combination of these features define the 7 basic classes of facial expression we want to classify (the Neutral class is also considered in classification).



(a) (b)

Figure 1: (a) The wireframe model, (b) the facial motion measurements

3. FINDING DEPENDENCIES AMONG FEATURES

The NB classifier was successful in many applications. However, the strong independence assumptions seem to be unreasonable in our case. It could be beneficial to search for another structure that captures better the dependencies among the features. Of course, the attempt to find all the dependencies is an NP-complete problem. So, we restrict ourselves to a smaller class of structures called the Tree-Augmented-Naive Bayes (TAN) classifiers. TAN classifiers have been introduced by Friedman, et al. [6] and are represented as Bayesian networks. Bayesian networks are acyclic graphical models, with the class and features as the nodes, and the dependencies represented by the directed edges in the graph between the nodes. 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 the class node and at most one other feature as parents, such that the result is a tree structure for the features. An example of a TAN classifier is given in Figure 2(a). Friedman, et al. [6] proposed the use of TAN model as a classifier, to enhance the performance over the simple Naive-Bayes classifier. TAN models are more complicated than the Naive-Bayes, but are not fully connected graphs. 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.

Learning the TAN classifier is more complicated. In this case, we do not fix the structure of the Bayesian network, but try to find the TAN structure that maximizes the likelihood function given the training data out of all possible TAN structures. In general, searching for the best structure has no efficient solution, however, searching for the best TAN structure does have one. The method is using the modified Chow-Liu algorithm [3] for constructing Tree-Augmented Bayesian networks [6]. This is done as follows:

1. Compute the class conditional pair-wise mutual information between each pair of features,

$$I_P(X_i, X_j|C) = \sum_{X_i, X_j, C} P(x_i, x_j, c) log \frac{P(x_i, x_j|c)}{P(x_i|c)P(x_j|c)}, i \neq j \quad (1)$$

2. Build a complete undirected graph in which each vertex is a variable and the weight of each edge is the mutual information computed in step 1.
3. Build a maximum weighted spanning tree (MWST).
4. Transform the undirected MWST of step 3 to a directed graph by choosing a root node and pointing the arrows of all edges away from the root.
5. Make the class node the parent of all the feature nodes in the directed graph of step 4.

This procedure ensures to find the TAN model that maximizes the likelihood of the data we have. The algorithm is computed in polynomial time ($O(n^2 log N)$, with N being the number of instances and n the number of features).

The learning algorithm for the TAN classifier is only feasible in cases where all the features are discrete. In our problem the features are continuous. The number of parameters of the TAN model grows exponentially with respect to the number of discrete values each feature takes. To solve this problem we propose a hybrid TAN and Gaussian classifier. We first discretize the features and use the TAN model learning algorithm to learn the dependency structure among the features. Then, we revert back to the original continuous features and model them as Gaussian, using the TAN graph structure. The added complexity of the Gaussian model is only linear in the number of features, but we are still able to capture dependencies among the features.

The full joint distribution of the Gaussian-TAN model can be written as:

$$p(c, x_1, x_2, ..., x_n) = p(c) \prod_{i=1}^{n} p(x_i|pa_{x_i}, c), \quad (2)$$

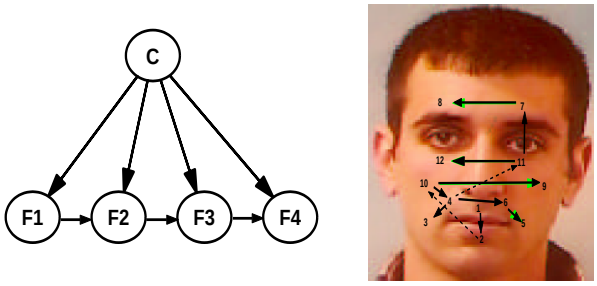
where pa_{x_i} is the feature that is the additional parent of feature x_i (empty for the root feature in the directed tree graph of step 4).

Using the Gaussian assumption, the pdf's of the distribution in the product above are:

$$p(X_i = x_i|pa_{x_i}, C = c) = N_c(\mu_{x_i} + a \cdot pa_{x_i}, \sigma^2_{x_i}(1 - \rho^2)) \quad (3)$$

where $N_c(\mu, \sigma^2)$ refers to the Gaussian distribution with mean μ and variance σ^2 given that the class is c, $\rho = \frac{COV(x_i, pa_{x_i})}{\sigma_{x_i}\sigma_{pa_{x_i}}}$ is the correlation coefficient between x_i and pa_{x_i}, and $a = \frac{COV(x_i, pa_{x_i})}{\sigma^2_{pa_{x_i}}}$. Estimating the Gaussian-TAN model involves estimating all the class conditional means and variances for each feature as in the NB model, then estimate the class conditional covariances between features and their feature parents. In terms of model complexity, there are $|C| \cdot (n - 1)$ extra parameters to estimate (the covariances).

Figure 2(b) shows the tree structure of the features learned using a database of subjects displaying different facial expressions. The arrows are from parents to children features (dashed lines represent links that are relatively weaker than the others). From the tree structure we see that the bottom half of the face is almost disjoint from the top portion, except for a weak link between AU 4 and AU 11.



(a) (b)

Figure 2: (a) An example of a TAN classifier. (b) The learned TAN structure for the facial features.

4. EXPERIMENTS

In order to test the algorithms described in the previous sections, we collected data of people that are instructed to display facial expressions corresponding to the six types of emotions. The data collection method is described in detail in [1]. 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, and always coming back to a neutral state between each emotion sequence. The video was used as the input to the face tracking algorithm described in Section 2. The sampling rate was 30 Hz, and a typical emotion sequence is about 70 samples long ($\sim$ 2 s).

The data was collected in an open recording scenario, where the person was asked to display the expression corresponding to the emotion being induced. This is of course not the ideal way of collecting emotion data. The ideal way would be using a hidden recording, inducing the emotion through events in the normal environment of the subject, not in a studio. The main problem with

collecting the data this way is the impracticality of it and the ethical issue of hidden recording.

We used the database described above to test our algorithms. We performed 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. For the TAN classifiers we used the dependencies shown in Figure 2(b), learned using the algorithm described in Section 3. As benchmark we also present the results obtained with a NB classifier when the underlying distribution assumption was Cauchy and Gaussian [11].

4.1. Person-Dependent Tests

There are six sequences of each facial expression for each person. For each test, one sequence of each emotion is left out, and the rest are used as the training sequences. Table 1 shows the recognition rate for each person and the total recognition rate averaged over the five people. It can be seen that taking into account the dependencies in the features (the TAN model) gives significantly improved results. It is also worth noting that the results for subject 5 are consistently worse for all classifiers. 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. The NB-Cauchy assumption does not give a significant improvement in recognition rate comparing with the NB-Gaussian assumption mainly due to the fact that in this case there are not many outliers in the data (each person was displaying the emotion sequences in the same environment). This may not be the case in a natural setting experiment.

Subject	NB-Gaussian	NB-Cauchy	TAN
1	80.97%	81.69%	85.94%
2	87.09%	84.54%	89.39%
3	82.5%	83.05%	86.58%
4	77.18%	79.25%	82.84%
5	69.06%	71.74%	71.78%
Average	79.36%	80.05%	83.31%

Table 1: Person-dependent expression recognition accuracies

In average the best results are obtained by using the TAN classifier, followed by the NB-Cauchy and the NB-Gaussian classifiers. The confusion matrix for the TAN classifier is presented in Table 2. The analysis of the confusion between different emotions shows that Happy and Surprise are well recognized. The other more subtle emotions are confused with each other more frequently, with Sad being the most confused emotion. These observations suggest that we can see the facial expression recognition problem from a slightly different perspective. Suppose that now we only want to detect whether the person is in a good mood, bad mood, or is just surprised (this is separated since it can belong to both positive and negative facial expressions). This means that we consider now only 4 classes in the classification: Neutral, Positive, Negative, and Surprise. Anger, Disgust, Fear, and Sad will count for the Negative class while Happy will count for the Positive class.

The confusion matrix obtained for the TAN classifier is presented in Table 3. The system can tell now with 87-92% accuracy if a person displays a negative or a positive facial expression.

Emotion	Neutral	Positive	Negative	Surprise
Neutral	79.58	1.21	15.92	3.29
Positive	1.06	87.55	8.65	2.74
Negative	6.48	0.1	92.15	1.26
Surprise	3.17	0.8	9.81	86.22

Table 3: Person-dependent average confusion matrix using the TAN assumption

4.2. Person-Independent Tests

In the previous section it was seen 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 subjects are used as training sequences. This test is repeated five times, each time leaving a different person out (leave one out cross validation). The recognition rates are 58.94% for NB-Gaussian, 63.50% for NB-Cauchy, and 65.11% for TAN. In this case they are lower compared with the person-dependent results meaning that the confusions between subjects are larger than those within the same subject.

In this case the TAN classifier also provides the best results. It is important to observe that the Cauchy assumption also yields an improvement compared to the NB-Gaussian classifier, due to the capability of the Cauchy distribution to handle outliers. 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 [5], 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 classifiers. Again we see that Surprise and Happy are detected with high accuracy, and other expressions are greatly confused.

If we now consider the problem where only the person mood is important, the classification rates are significantly higher. The confusion matrix obtained for the TAN classifier is presented in Table 5. Now the recognition rates are much higher. The system can tell now with about 77% accuracy if a person displays a negative or a positive facial expression.

Emotion	Neutral	Positive	Negative	Surprise
Neutral	76.95	0.46	21.09	1.50
Positive	3.21	77.34	15.48	3.97
Negative	11.25	4.45	77.57	6.73
Surprise	4.97	6.83	14.09	74.11

Table 5: Person-independent average confusion matrix using the TAN assumption

5. DISCUSSION

In this work we presented a method for expression recognition from video. We introduced a TAN classifier that learns the dependencies between the facial features and we presented an algorithm that finds the best TAN structure. Using this structure we obtained significantly improved results over the simple NB models when person-dependent and person-independent tests were conducted. Moreover, we showed that when the facial expression recognition prob-

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	3.17	0.80	0.52	2.45	5.73	1.08	86.22

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

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	6.41	2.95	4.38	74.11

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

lem is reduced to a mood recognition problem the classification results are significantly higher.

One of the main drawbacks in all of the works done on emotion recognition from facial expression videos is the lack of a benchmark database to test different algorithms. This work relied on a database collected by Chen [1], but it is difficult to compare the results to other works using different databases. The recently constructed database by Kanade et al [8] will be a useful tool for testing these algorithms.

Are the recognition rates sufficient for real world use? We think that it depends upon the particular application. In the case of image and video retrieval from large databases, the current recognition rates could aid in finding the right image or video by giving additional options for the queries. For future research, the integration of multiple modalities such as voice analysis and context would be expected to improve the recognition rates and eventually improve the computer's understanding of human emotional states. Voice and gestures are widely believed to play an important role as well, and physiological states such as heart beat and skin conductivity are being suggested. People also use context as an indicator of the emotional state of a person. This work is just another step on the way toward achieving the goal of building more effective computers that can serve us better.

Acknowledgments. We would like to thank Prof. Fabio G. Cozman for the use of his code implementing the Naive Bayes and TAN classifiers in the Java language, using the libraries of the JavaBayes system (available at http://www.cs.cmu.edu/~javabayes). This work has been supported in part by the National Science Foundation Grants CDA-96-24396 and IIS-00-85980. The work of Ira Cohen and Asutosh Garg is supported by Hewlett Packard and IBM fellowships, respectively.

6. REFERENCES

[1] L. S. Chen. *Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction*. PhD thesis, University of Illinois at Urbana-Champaign, 2000.

[2] L. S. Chen, H. Tao, T. S. Huang, T. Miyasato, and R. Nakatsu. Emotion recognition from audiovisual information. In *IEEE Workshop on Multimedia Sign. Proces.*, pages 83–88, 1998.

[3] C.K. Chow and C.N. Liu. Approximating discrete probability distribution with dependence trees. *IEEE Trans. Inf. Theory*, 14:462–467, 1968.

[4] L. C. De Silva, T. Miyasato, and R. Natatsu. Facial emotion recognition using multimodal information. In *IEEE Int. Conf. Inf. Communic Sign. Proces.*, pages 397–401, 1997.

[5] P. Ekman and W. V. Friesen. *Facial Action Coding System: Investigator's Guide*. Consulting Psychologists Press, 1978.

[6] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2):131–163, 1997.

[7] D. Goleman. *Emotional Intelligence*. Bantam Books, 1995.

[8] T. Kanade, J.F. Cohn, and Y. Tian. Comprehesive database for facial expression analysis. In *Int. Conf. Automatic Face Gesture Rec.*, pages 46–53, 2000.

[9] K. Mase. Recognition of facial expression from optical flow. *IEICE Transactions*, E74(10):3474–3483, 1991.

[10] M. Rosenblum, Y. Yacoob, and L.S. Davis. Human expression recognition from motion using a radial basis function network architecture. *IEEE Trans. Neural Network*, 7(5):1121–1138, 1996.

[11] N. Sebe, I. Cohen, A. Garg, M. Lew, and T. Huang. Emotion recognition using a Cauchy Naive Bayes classifier. In *ICPR*, 2002, to appear.

[12] H. Tao and T. S. Huang. Connected vibrations: A modal analysis approach to non-rigid motion tracking. In *CVPR*, 1998.

[13] Y. Yacoob and L.S. Davis. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Trans. Patt. Anal. Machine Intell.*, 18(6):636–642, 1996.