MACHINE LEARNING TECHNIQUES FOR FACE ANALYSIS

D. Datcu and L.J.M. Rothkrantz
Man Machine Interaction Group
Delft University of Technology
Mekelweg 4, 2628 CD Delft
The Netherlands
E-mail: {D.Datcu, L.J.M.Rothkrantz}@ewi.tudelft.nl

KEYWORDS
Machine learning, pattern recognition, classifiers, face detection, facial expression recognition.

ABSTRACT

Facial related analysis represented milestones in the fields of computer vision for many decades. Lots of methods have been designed and implemented so as to solve the specific requirements. In the current paper we present three different classification algorithms that we use to fulfill the tasks concerning face detection and facial expression recognition. One of the methods, Relevance Vector Machines (RVM) stands for a novel supervised learning technique that is based on a probabilistic approach of Support Vector Machines. The mathematical base of the models is presented. The data for testing were selected from the Cohn-Kanade Facial Expression Database. We report recognition rates for six universal expressions based on a range of experiments. Some discussions on the comparison of different classification methods are included.

INTRODUCTION

Human computer interaction stands for a major step towards making machines have an even more important role to play in the human life. It is based on interdisciplinary researches that aim to implement knowledge from behavioral and social sciences into machines. It is the human nature that we can estimate a person's psychological state following the observation on his face. Nonverbal communication channels are typically set during common interpersonal relations and visual messages are processed in a transparent manner. The general tendency is to construct robotic systems that are able to understand the environmental world and to interact with the existent actors. Human-computer interfaces play an essential role in the perception and feedback the system is capable of. In this context, the advantage of making machines to read human facial expressions is tremendous. Facial expressions reveal internal characteristics of the expresser. To address the problem of facial expression recognition, in our approach we extract parametric information with high discrimination power from facial feature space and use it in a data-driven classification environment. The current paper primarily focuses on the aspects related to the classification methods for facial expression recognition. Secondly, techniques related to vision have to be involved for processing the video signal for detection of faces.
The classifiers are aimed at solving the universal problem of classification. We begin from Support Vector Machines SVM (Vapnik 1995) that is based on a solid mathematical foundation. The Naive Bayes classifier (Langley et al. 1992) is also introduced and practical aspects on the performance are presented. Then the novel Relevance Vector Machines RVM (Tipping 2000) is introduced as an alternative to SVM. The difficulty of the automatic analysis of facial expressions (Pantic and Rothkrantz 2000) resides in the variety of characteristic appearance with respect to both individuality and face anatomic dynamics. The inner complexity makes from the processes of feature detection and feature oriented expression recognition difficult tasks. To our knowledge, this is the first research that involves Relevance Vector Machines for facial expression recognition.

RELATED WORK

The recognition of facial expressions implies finding solutions to three distinct types of problems. The first one relates to detection of faces in the image. Once the face location is known, the second problem is the detection of the salient features within the facial areas. The final analysis consists in using any classification model and the extracted facial features for identifying the correct facial expression. For each of the processing steps described, there have been developed lots of methods to tackle the issues and specific requirements. Depending on the method used, the facial feature detection stage involves global or local analysis.
The internal representation of the human face can be either 2D or 3D. In the case of global analysis, the connection with certain facial expressions is made through features determined by processing the entire face.
The efficiency of methods as Artificial Neural Networks or Principal Component Analysis is greatly affected by head rotation and special procedures are needed to compensate the effects of that. On the other hand, local analysis performs encoding of some specific feature points and use them for recognition. The method is actually used in the current paper. However, other approaches have been also used at this layer.
One method for the analysis is the internal representation of facial expressions based on collections of Action Units (AU) as defined in Facial Action Coding System (FACS) (Bartlett et al. 2004; Ekman and Friesen 1978) . It is one of the most efficient and commonly used methodology to handle facial expressions. Some attempts to automatically detect the salient facial features implied computing descriptors such as

The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.

scale-normalized Gaussian derivatives at each pixel of the facial image and performing some linear-combinations on their values. It was found that a single cluster of Gaussian derivative responses leads to a high robustness of detection given the pose, illumination and identity (Gourier et al. 2004). A representation based on topological labels is proposed in (Yin et al. 2004). It assumes that the facial expression is dependent on the change of facial texture and that its variation is reflected by the modification of the facial topographical deformation. The classification is done by comparing facial features with those of the neutral face in terms of the topographic facial surface and the expressive regions.

A robust face detection technique was developed in (Viola and Jones 2001) based on a cascading classifier that include a set of so called 'weak' classifiers. The features stand for values of difference between the sums of pixel intensities computed in different areas in the image. Some approaches firstly model the facial features and then use the parameters as data for further analysis such as expression recognition. The system proposed by (Moriyama et al. 2004) is based on a 2D generative eye model that implements encoding of the motion and fine structures of the eye and is used for tracking the eye motion in a sequence. As concerning the classification methods, various algorithms have been developed (Pantic and Rothkrantz 2000), adapted and used during time. Neural networks have been used for face detection and facial expression recognition (Stathopoulou and Tsihrintzis 2004; deJong and Rothkrantz 2004). The second reference directs to a system called Facial Expression Dictionary (FED) (deJong and Rothkrantz 2004) that was a first attempt to create an online nonverbal dictionary. Other classifiers included Bayesian Belief Networks (BBN) (Datcu and Rothkrantz 2004), Expert Systems (Pantic and Rothkrantz 2000) or Support Vector Machines (SVM) (Bartlett et al. 2004). Other approaches have been oriented on the analysis of data gathered from distinct multi-modal channels. They combined multiple methods for processing and applied fusion techniques to get to the recognition stage (Fox and Reilly 2004).

VISUAL FEATURE MODEL

Although local analysis is sensitive to identity and partial occlusions, we overcome that by increasing the variability of data for training the model and by increasing the parameter redundancy. The variability is handled by the face characteristics in the Cohn-Kanade database (Kanade et al. 2000). The redundancy assumes the use of feature parameters on an asymmetric model, i.e. in case of occlusion or low visibility for one eye, the recognition is done by taking into account the data from the other eye. The task prior to classification stage is aimed at preparing the feature data associated with the input face. Depending on the type of the classifier involved in the next step, the data have to be transformed to a certain format.

In the current approach a transform $\Gamma: \theta \to \vartheta$ converts the facial feature image θ to some parameters $p_i \in \vartheta, i = 1,...,L$ of an intermediate model. The parameterization of the facial features has the advantage of providing the classifier with data that encode the most important aspects of the facial expressions. Furthermore, it acts as a dimensionality reduction procedure since the dimension of the feature space is lower than the dimension of the image space.

An advantage of the model is that it also can handle a certain degree of asymmetry by using some parameters for both left and right sides of the face. Each facial feature $\kappa \in \theta$, θ={Left/Right eye, Left/Right Eyebrow, Mouth, Chin} is extracted at the previous processing stage by distinct processing channels.

The transform Γ firstly extracts the location of each FCP from the input facial feature κ. Eventually the feature parameters p_i are computed as the values of some angles and/or Euclidean distances between key points assumed to reflect the location of the facial features.

The symmetry of the model is assumed to make the recognition process of facial expressions robust to occlusion or poor illumination i.e. if the left eye area is not directly visible do not use related information.

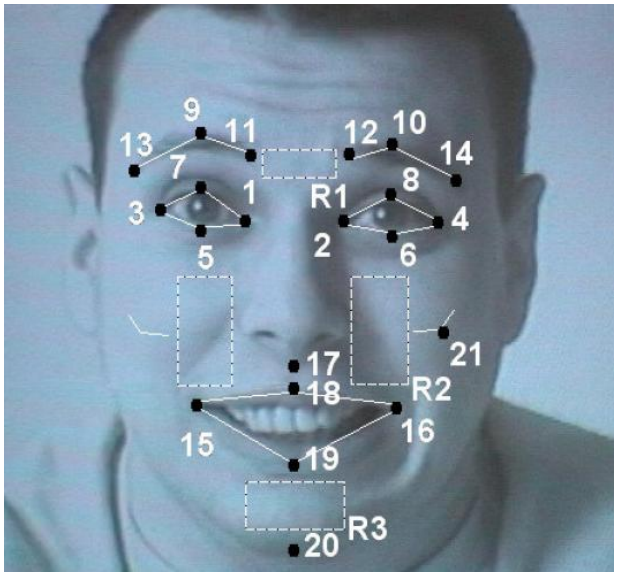

Figure 1: FCP set

The key points are defined as Facial Characteristic Points (FCPs) and the FCP-set (Figure 1) is based on an extension of Kobayashi & Hara model (Kobayashi and Hara 1972).

The final step of preprocessing was related to scale all the distances so as to be invariant to the size of the image. For the face detection, a procedure is run on each frame in the video sequence. First a 19 × 19 sliding window is defined. The processing for one window implies different tasks, according to the model used. If the image bitmap is used as an unique feature, then the set of pixel intensities is provided directly to the face/non-face classifier. Otherwise additional processing are required i.e. basic operations on pixel intensities (Viola and Jones 2001).

The cases when the dimension of an existent face is different than that of the window, a multiresolution pyramid is computed and analysis take place on each level. The process ends with the list of face locations in the current frame. An alternative consists in using feature detection and based on the findings to determine the location of any faces.

CLASSIFIERS

The current section is aimed at presenting the theoretical background of the classification techniques used in the research.

If X denotes the space of input variables representing the face images and T the space of output variables i.e. the facial expression label, then f is the associated deterministic function and $t_n = f(x_n) + \varepsilon_n$ represents the possible overlapping target values. The training database is:

$$Z = \{(x_i, t_i) \in X \times T \mid i = 1,..., M\}$$

The learning step implies the use of the training database Z together with any other prior knowledge for finding a function $\hat{f}$ out of a class of functions F, that encodes the estimated dependency. The process can be seen as a transform Ψ of the initial data v into some internal knowledge ν of the interest phenomena and so $\Psi : v \to \nu$. From the notations used above, $Z \subseteq v$ and $\hat{f} \subseteq \nu$. The result function $\hat{f}$ is assumed to produce an efficient classification of the facial expression label t given a new feature vector x.

Support Vector Machine (SVM)

The Support Vector Machine (SVM) algorithm has been successfully used in classification related problems since it was introduced by Vapnik (Vapnik 1995) in the late 1970's. The idea was that given the collection of input-target pairs Z with $X \in R^n$ and $T = \{-1,+1\}$, a hyperplane $f \in F_h$, $F_h = \{f : X \to R \mid (w^T x) + b\}$ with the maximal margin has to be found as a solution of an optimization problem. The distribution of the two classes is such that they are linearly separable. The constraints aim at determining the model parameters $\{w, b\}$ that fit the training data and minimize the complexity of the decision function in the same time. The result is a classifier with a certain level of robustness to over-fitting. The margin represents a measure of class separation efficiency and is defined as the Euclidean distance between the data and the separating hyperplane.

Non-linearity that is specific to facial expression representation is handled through kernel methods (non-linear SVM) that first preprocess the data by non-linear mapping $\phi : R^N \to \varepsilon$ and then apply the linear algorithm in the image space ε. The image space is a vector space of functions $\varepsilon = \{f \mid f : X \to R\}$. The positive definite kernel function $k : X \times X \to \phi$ acts as a dot-product over ϕ and the mapping is expressed as $\phi(x) = k(\cdot, x)$.

Naive Bayesian Classifier (NB)

The Naive Bayesian (NB) is a probabilistic classifier based on the assumption that the set of model parameters are independent given the class parameter. An example E is presented as a tuple of attribute feature values $\langle x_1, x_2,..., x_n \rangle$. The parameter C represents the classification variable and c_j is a value of C. The probability of a sample being class c_j is:

$$p(c_j \mid E) = \frac{p(E \mid c_j) p(c_j)}{p(E)}$$

The Bayesian approach to classifying a new instance relates to assigning the most probable target value, given the attribute values $\langle x_1, x_2,..., x_n \rangle$ describing the instance.

$$\begin{aligned} f_b(E) &= \arg\max_{c_j} \quad p(c_j \mid x_1, x_2,..., x_n) \\ &\cong \arg\max_{c_j} \quad p(c_j \mid x_1, x_2,..., x_n) p(c_j) \end{aligned}$$

where $f_b(E)$ is called a Bayesian classifier. Considering that all the attributes are independent given the value of the class variable, then:

$$p(E \mid c_j) = p(x_1, x_2,..., x_n \mid c_j) = \prod_{i=1}^{n} p(x_i \mid c_j)$$

By rewriting, the previous formula of the classifier becomes:

$$p(E \mid c_j) = \arg\max_{c_j} \quad p(c_j) \prod_{i=1}^{n} p(x_i \mid c_j)$$

The function $f_{nb}(E)$ is called the Naive Bayesian classifier or naive Bayes (NB). Naive Bayesian is the simplest case of Bayesian network. The assumption of conditional independence that is the independence of attribute variables given the class variable is not properly fulfilled for the real case. For facial expression recognition, some of the parameters $p_i \in \vartheta, i = 1,..., L$ do not follow the independence assumption because they measure the behavior of related components of the same facial features. Moreover, the parametric changes of different facial features are correlated due to the dynamics of facial expressions. Though, the naive Bayes classifier performs very well even in these cases. An explanation given for the performance of the binary NB classifier is presented in (Zhang 2004).

Relevance Vector Machines (RVM)

Tipping introduced the Relevance Vector Machine (RVM) (Tipping 2000) as a probabilistic sparse kernel model based on the support vector machine theory. Each of the model's weights has associated a prior that is characterized by a set of hyperparameters whose values are determined during the learning process. Following the above notations, $p(y \mid x)$ is assumed to be Gaussian $N(y \mid f(x), \sigma^2)$ and the mean of the distribution is computed as specified for the SVM. The overfitting effect that occurs while determining the parameter values of the likelihood $p(y \mid w, \sigma^2)$ is confined by including an ARD Gaussian prior over the weights w_i as

$$p(w \mid \alpha) = \prod_{i=0}^{N} N(w_i \mid 0, \alpha_i^{-1})$$

The dataset likelihood is expressed by using logistic functions, in the form:

$$p(w \mid w) = \prod_{n=1}^{N} \sigma\{f(x_n)\}^{y_n} [1 - \sigma\{f(x_n)\}]^{1-y_n}$$

The parameters α_i and w_i of the model are computed through an iteration procedure until convergence is achieved. One advantage over SVM is that for comparable generalization performance, it uses fewer kernel functions. This determines less memory and time for processing and so making possible for the usage of RVM as a real-time classifier. By using RVM, the relevance vectors stand for representative training samples of the emotional classes rather than data points closer to the separation hyperplane as in SVM model.

EXPERIMENTS

For testing the face detection by using different classifiers, the CBCL (the Center for Biological and Computational Learning, MIT) face database was used. It contains 19×19 image samples. The training set contains 2429 faces and 4548 non-faces and the testing set contains 472 faces, 23573 non-faces. The classification process implied the use of all the image map as a unique feature. Testing face detection with RVM classifier for 1000 black-white samples in the database and using 5-fold cross validation, we achieved a detection rate of 99.20%. The number of relevance vectors was 392.

Ongoing work aims at using the method presented in (Viola and Jones 2001) for determining a set of representative features to be used for classification. More exactly, each feature would stand for one parameter computed as a difference between the sums of pixel intensities over some interest areas. The idea is to have a set of such features that together would help for discriminating between a face or non-face.

Table 1: Recognition mismatch rate for facial expression recognition by using SVM classifier

Expression	Mismatch rate	Number of support vectors
Surprise	14.32 ± 1.80%	63
Sadness	3.06 ± 2.78%	23
Anger	5.91 ± 1.55%	46
Happy	3.16 ± 2.47%	38
Disgust	9.54 ± 1.97%	34
Fear	24.97 ± 2.07%	72

An initial process of automatically generating features is run in the image space. Then AdaBoost (Adaptive Boosting) technique (Freund and Schapire 1995) is used to select only the most expressive features from the initial generated set. The final 'strong' classifier is expressed as a sequence of 'weak' classifiers, each classifier relying on only one certain feature from the selected set. The final cascading detector is created using a weighted voting scheme so as the weight of each classifier depends on its performance on the training set.

The data used for training the models for the facial expression recognition experiments have been processed from the Cohn-Kanade database. The database contains approximately 2000 image sequences from 200 subjects ranged in age from 18 to 30 years. Sixty-five percent were female, 15 percent were African-American and three percent were Asian or Latino.

In the first step, only 485 images have been used for the experiments, each image representing a sample data for the model. Then the classification was improved by adding new samples. In the case of Naive Bayes, the error dropped from 33% to 26.89%.

Table 2: Recognition mismatch rate for facial expression recognition by using Naive Bayes classifier

Expression	Mismatch rate
Surprise	11.66 ± 7.29%
Sadness	22.35 ± 5.98%
Anger	31.01 ± 12.60%
Happy	17.76 ± 7.79%
Disgust	38.82 ± 12.85%
Fear	39.75 ± 16.95%

Some specific steps were passed to extract and prepare the data to comply with the requirements of the classification process. The algorithms are set to perform classification on the 6 universal expressions (Happy, Anger, Sad, Surprise, Disgust, Fear). The classification results of the facial expression recognition can be analyzed by looking at the mismatch rate. The method used for computing the error was leave-5-out cross validation. The distribution of samples in the database for testing was as follows:
N(Surprise, Sadness, Anger,Happy,Disgust, Fear) = (108, 92, 30, 110, 59, 86).

Table 3: Recognition mismatch rate for facial expression recognition by using RVM classifier

Expression	Mismatch rate	Number of relevance vectors
Surprise	6.25 ± 1.51%	21
Sadness	12.29 ± 2.57%	34
Anger	5.00 ± 1.87%	15
Happy	7.92 ± 1.71%	23
Disgust	8.54 ± 1.91%	25
Fear	15.00 ± 2.38%	38

As it can be seen (Table 1 and Table 3), the error rate in the case of RVM (9.16%) is comparable to that of SVM (10.15%) classifier. One important aspect is that in case of RVM classifier the number of relevance vectors (276) is greater than the number of support vectors (156) of SVM. The effect is a decrease of the number of kernel functions and so of the complexity of the model. In terms of practical characteristics that means less processing time and also less memory for using this type of classifier.

The Naive Bayes classifier stands for an easy way to predict the class of an unseen sample by means of integrating probability knowledge computed from the known examples.

However, the performance of NB is based on the training data. The training stage implies the computation of the dependencies of each feature given the sample class. From this, the system learns about the dependency distributions of the data. If the training set does not contain enough data for learning, the result would be an approximation of the distributions that does not exactly fit the real ones. The degree of error on the real distributions greatly affects the classification results. For testing NB classifier for facial expression recognition, the cross fold validation procedure was used, where the number of folds was 10. It also implied an increased number of samples as N(Surprise, Sadness, Anger,Happy,Disgust, Fear) = (108, 206, 105, 110, 108, 86). Because of that, the comparison with the other techniques may be improper. The facial recognition general mismatch rate for NB is 26.89% (Table 2). Nevertheless, the analysis of facial expressions in static images has its own limitations. That can be mainly explained through the dynamics characteristics of the salient features involved in facial expressions' structure. An important improvement for the recognition system may include also the encoding and usage of the knowledge over these elements (Datcu and Rothkrantz 2004).

CONCLUSION

Facial expression recognition has scientifically been considered a real challenging problem in the fields of pattern recognition or robotic vision. The current research aims at proposing Relevance Vector Machines (RVM) as a novel classification technique for the recognition of facial expressions in static images. The results presented highlight the potential of the Relevance Vector Machines as a facial expression classifier and for face detection. The exemplifications start from the idea of the Support Vector Machines and addresses the issues concerning the use of two types of classifiers in the context of facial expression recognition problem. The RVM is a relatively new classification method and this work is the first one that uses the technique as a recognition engine for facial expressions. The fundamental aspects are described on both theoretical and practical sides. Each classifier model presents certain advantages and limitations and have been designed so as to perform prediction on the static images. The results for RVM show that it is suitable for facial expression classification in static images and it leads to a decrease of complexity comparing to SVM. The still image analysis is very restrictive with respect to the subtle dynamics of the facial features.

Additional research has been conducted to encode temporal behavior in the classification models so as to make possible the use of the recognition systems to run on image sequences. Another idea for increasing the capabilities and efficiency is to make use of fusion techniques to handle multiple modalities.

REFERENCES

Bartlett, M.S.; G.Littlewort; C.Lainscsek; I.Fasel; and J.Movellan. 2004. "Machine learning methods for fully automatic recognition of facial expressions and facial actions". Proceedings of IEEE SMC. 592–597.

Datcu, D.; and L.J.M.Rothkrantz. 2004. "Automatic recognition of facial expressions using bayesian belief networks". Proceedings of IEEE SMC. 2209–2214.

Ekman, P.; and W.V.Friesen. 1978. "Facial action coding system:investigator's guide". Consulting Psychologists Press.

Fox, N.A.; and R.B.Reilly. 2004. "Robust multi-modal person indentification with tolerance of facial expression". Proceedings of IEEE SMC. 580–585.

Freund Y.; and R.E.Schapire.1995. "A decision theoretic generalization of on-line learning and an application of boosting". Proceedings of the Second European Conference on Computational Learning Theory. Springer-Verlag. 23–37.

Gourier, N.; D.Hall; and J.L.Crowley. 2004. "Facial feature detection robust to pose, illumination and identity". Proceedings of IEEE SMC. 617–622.

Kanade, T.; J.Cohn; and Y.Tian. 2000. "Comprehensive database for facial expression analysis". Proc. IEEE Int'l Conf. Face and Gesture Recognition. 46–53.

Kobayashi, H.; and F.Hara. 1972. "Recognition of mixed facial expressions by neural network". IEEE International workshop on Robot and Human Communication. 381–386.

Langley, P.; W.Iba; and K.Thompson. 1992. "An analysis of bayesian classifiers". Proceedings of Tenth National Conference on Artificial Intelligence. AAAI Press and MIT Press. 223–228.

Moriyama, T.; J.Xiao; J.F.Cohn; and T.Kanade. 2004. "Meticulouslydetailed eye model and its application to analysis of facial image". Proceedings of IEEE SMC. 580–585.

Pantic, M.; L.J.M.Rothkrantz. 2000. "Self-adaptive expert system for facial expression analysis". Proceedings of IEEE SMC. 73–79.

Pantic, M.; L.J.M.Rothkrantz. 2000. "Automatic analysis of facial expressions: the state of the art". IEEE Trans. PAMI. 22(12).

Stathopoulou, I.O.; and G.A.Tsihrintzis. 2004. "An improved neural-network-based face detection and facial expression classification system". Proceedings of IEEE SMC. 666–671. deJong, E.J.; and L.J.M.Rothkrantz. 2004. "Fed - an online facial expression dictionary". Proceedings of Euromedia. 115–119.

Tipping, M. E. 2000. "The relevance vector machine". In Sara A Solla, Todd K Leen, and Klaus-Robert Müller, editors. Advances in Neural Information Processing Systems 12. Cambridge. Mass: MIT Press.

Vapnik, V. 1995. "The nature of statistical learning". Springer, New York.

Viola, P.; and M.Jones. 2001. "Robust real-time object detection". Workshop on Statistical and Computational Theories of Vision-Modeling Learning, Computing and Sampling.

Yin, L.; J.Loi; and W.Xiong. 2004 "Facial expression analysis based on enhanced texture and topographical structure". Proceedings of IEEE SMC. 586–591.

Zhang, H. 2004. "The optimality of naive bayesian".