Relevant Component Analysis for static facial expression classification

Abstract

This paper addresses the issue of automatic classification of the six universal emotional categories (joy, surprise, fear, anger, disgust, sadness) in the case of static images. Appearance parameters are extracted by an active appearance model(AAM) representing the input for the classification step. We introduce Relevant Component Analysis (RCA) in the context of facial expression recognition framework and we test this method against several other classification techniques, including LDA, GDA and SVM, on the Cohn-Kanade database.

1. Introduction

In the recent years there has been an increasing interest in computational facial expression analysis, above all as a way to achieve an effective natural human-machine interaction. Facial expressions are one of the most powerful means to convey emotions and governing the way we relate to each other. Indeed Blum [5] states that "The face is the most extraordinary communicator, capable of accurately signaling emotion in a bare blink of a second, capable of concealing emotion equally well", while Darwin [9] already underlined in 1872 the universality of facial expressions. In 1971 Ekman and Friesen [11] studied facial expressions in several disparate cultures, mapping the most minute twitches in thousands of expressions. From these, they distilled the six primary emotions carrying each a distinctive content, together with a unique facial expression. The six universally recognized facial expressions are : happiness, sadness, fear, anger, disgust and surprise [12]. Thanks to advances in images processing, machine learning and pattern recognition, automatic facial expression recognition has become an active research topic in the statistical learning community. All the automatic facial expression recognition systems developed in the recent years share the same structure: they first extract features, then these facial features are used as inputs of a classification system, giving one of the preselected facial emotions as outcome. A proper face detection is fundamental in order to achieve good recognition performance. Facial features extraction methods can be categorized according to whether they focus on motion or deformation of faces. Lien et al. [15] analyzed holistic face motion with the aid of wavelet-based, multi-resolution dense optical flow, while Mase and Pentland [16] use a region based optical flow in order to estimate the activity of 12 of the 44 facial muscles. Gabor wavelet based filters have been largely used [3, 18] to detect line and edge borders over multiple scales and different orientations. Active Appearance Models (AAM) have been successfully used for face representation and relevant information extraction [8, 20, 1, 14]. AAM is the feature extractor method we decided to use in our work. This technique elegantly combines shape and texture models, in a statistical-based framework. Statistical analysis is performed through consecutive PCAs respectively on shape, texture and the combination of both. The combined model allows the AAM to have simultaneous control of shape and texture by a single vector of parameters representing our features. Details on AAM will be part of the subjects tackled in Section 2. Once a proper face representation has been defined, the recognition step will decide to which class the represented face belongs to. Facial expression analysis can be preformed from static images [7, 1] or video sequences [7, 6, 17]. Cohen et al.[7] introduced and tested different Bayesian network classifiers and a neural network approach. Abboud et al.[1] projected the AAM coefficients in the linear discriminant analysis(LDA) space and classified the tested image to the closest expression cluster. On his dynamic approach Cohen [7] proposed a multi-level hidden markov model classifier for automatically segmenting and recognizing human facial expression from video sequences. A manifold based dynamic approach for facial expression analysis has been recently proposed by Changbo et al. [6]. Changbo proposed a probabilistic expression classification method, integrating expression tracking and recognition in a cooperative system. We present here the use of Relevant Component Analysis(RCA) [19] as a metric learner in the task of expression classification only for static images. We will compare the results obtained with the RCA and the RCA combined with dimensionality reduction techniques to the ones using Abboud [1] approach and the ones given by some other linear and nonlinear classifiers. The remainder of this paper is organized as follows: in Section 2 we shortly review AAM and RCA. Section 3

describes the framework and the database used for the experiments which are reported in Section 4. Conclusions and future works are reported in Section 5.

2. Background overview

2.1. Active Facial Appearance Model

The AAM is a statistical-based method for matching a combined model of shape and texture to unseen faces. Statistical appearance models are generated by the combination of a model of shape variation with a model of texture variation. The setting-up of the model relies on a set of annotated images. The annotation consists of a group of landmark points around the main facial features, marked in each example. The shape is represented by a vector $\mathbf{s}$ brought into a common normalized frame -w.r.t. position, scale and rotation- to which all shapes are aligned. After having computed the mean shape $\bar{\mathbf{s}}$ and aligned all the shapes from the training set by means of a Procrustes transformation [10], it is possible to warp textures from the training set onto the mean shape $\bar{s}$, in order to obtain shape-free patches. Similarly to the shape, after computing the mean shape-free texture $\bar{g}$, all the textures in the training set can be normalized with respect to it by scaling and offset of luminance values. Eigen-analysis (PCA) is applied to build the statistical shape and textures models:

$$\mathbf{s_i} = \bar{\mathbf{s}} + \mathbf{\Phi_s b_{si}} \quad and \quad \mathbf{g_i} = \bar{\mathbf{g}} + \mathbf{\Phi_t b_{ti}} \tag{1}$$

where $\mathbf{s_i}$ and $\mathbf{g_i}$ are, respectively, the synthesized shape and shape-free texture, Φ_s and Φ_t are the matrices describing the modes of variation derived from the training set, b_{si} and b_{ti} the vectors controlling the synthesized shape and shape-free texture. The unification of the presented shape and texture models into one complete appearance model is obtained by concatenating the vectors b_{si} and b_{ti} and learning the correlations between them by means of a further PCA. The statistical model is then given by:

$$\mathbf{s_i} = \bar{\mathbf{s}} + \mathbf{Q_s c_i} \quad and \quad \mathbf{g_i} = \bar{\mathbf{g}} + \mathbf{Q_t c_i} \tag{2}$$

where Q_s and Q_t are the matrices describing the principal modes of the combined variations in the training set and $\mathbf{c_i}$ is the appearance parameters vector, allowing to control simultaneously both shape and texture. Fixing the parameters c_i we derive the shape and the shape-free texture vectors using equations (2). A full reconstruction is given by warping the generated texture into the generated shape. In order to allow pose displacement of the model, other parameters must be added to the appearance parameters c_i: the pose parameters p_i. The matching of the appearance model to a target face can be treated as an optimization problem, minimizing the difference between the synthesized model image and the target face [20].

2.2. Relevant Component Analysis Algorithm

Relevant Component Analysis (RCA) is a simple and efficient algorithm for learning a Mahalanobis distance. Many learning algorithms use a distance function over the input data as a principal tool and their performance critically depends on the quality of the metric. It follows that learning a good metric from the examples is an essential step to a successful application of these algorithms. RCA is a method that seeks to identify and down-scale global unwanted variability within the data. The method performs a projection of the input data into a feature space by means of a linear transformation which assigns a large weight to "relevant dimensions" and small weight to "irrelevant dimensions". The algorithm is based on the use of *chunklets*. A *chunklet* is a container of elements in equivalence relation among each others, meaning that they belong to the same although unknown class. The RCA aims to reveal the inherent structure of the data in the new feature space for that it can be used as a preprocessing step for unsupervised clustering or nearest neighbor classification.
The RCA procedure can be summerised as follow:

1. For each chunklet, subtract the chunklet's mean from all the points it contains and compute the within chunklet covariance matrix.

$$\hat{C} = \frac{1}{N}\sum_{j=1}^{n}\sum_{i=1}^{n_j}(\mathbf{x_{ji}} - \mathbf{m_j})(\mathbf{x_{ji}} - \mathbf{m_j})^t \tag{3}$$

   where $\mathbf{m_j}$ denotes the mean of the j-th chunklet and $\mathbf{x_{ji}}$ the i-th vector element of the j-th chunklet.

2. If needed apply dimensionality reduction to the data using $\hat{C}$ [19].

3. Compute the whitening transformation associated with $\hat{C}$: $W = \hat{C}^{-\frac{1}{2}}$ and apply it to the data points: $X_{new} = WX$, where X refers to the data points after dimensionality reduction, when applicable.

The whitening transformation plays an essential role in unraveling the structure of the data. W assigns lower weight to some directions in the original space; those are the directions in which the data variability is mainly due to within class variability, in other terms the "irrelevant" variability for the task of classification.

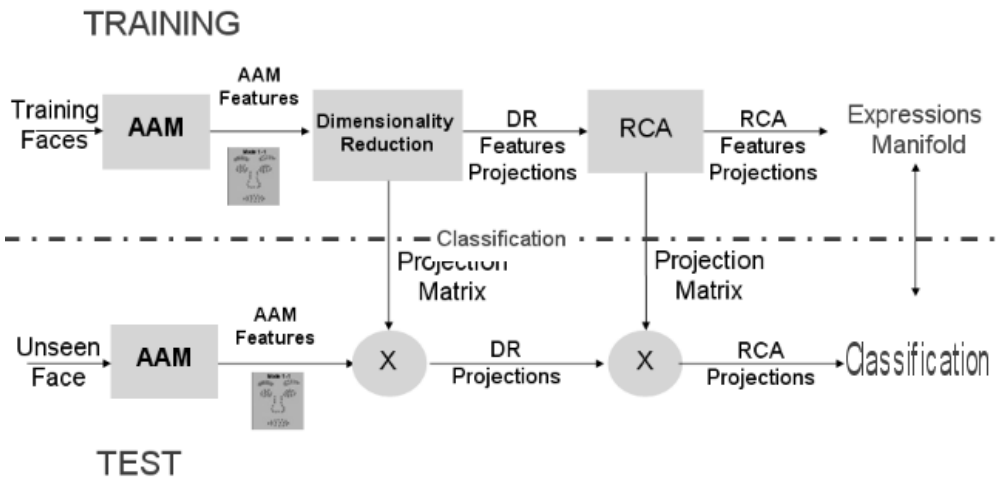

Figure 1. Our proposed facial expression recognition system

3. RCA for facial expression classification

The task here is to define and evaluate the performance of a facial expressions recognition system using a nearest neighbor classifier based on the RCA distance. We give a short overview on the classifiers we used to benchmark our proposed system.

The block-scheme of the framework is showed in Fig.1. In the scheme we consider the setting-up of an active appearance model as pre-processing step. The model has been built on a set of manually landmarked images. The upper half of Fig.1 represents the learning phase of the process: a training set of facial expressions (different from that of the AAM) is presented to our feature extractor. The appearance vector $\mathbf{c_i}$, corresponding to the matched appearance mask of the i-th training image, represents the feature on which our expression recognition system will rely on. The effective expression recognition training set is represented by the collected matrix C of appearance parameters. The goal of the remaining part of the training chain is to learn a discriminative manifold of expressions. Before applying RCA on the training data we perform a reduction in dimensionality.

As Bar-Hillel et al. states in the context of face recognition [2], RCA can be viewed as an augmentation of the standard, fully supervised Fisher's Linear Discriminant (FLD), which whitens the output of FLD w.r.t. the within class covariance. In the same paper the authors show how the use of FLD in combination with RCA dramatically improves the performance of the RCA. As mentioned in 2.2, the RCA algorithm requires the use of chunklets. There are two possible uses of chunklets, in the first one all data points are assigned to chunklets, while in the other only part of the data is assigned to chunklets. Clearly the fully supervised scheme gives better results then the partially labeled one. In our work we use the fully supervised scheme. At the output of the RCA block we obtain a new feature representation of the data space, the expressions manifold, in which Euclidean distance is less affected by irrelevant variability. It can be shown [19] that the nearest neighbor classification based on the Euclidean distance in the transformed space is statistically optimal.

In classification, an unseen face is presented to the feature extractor. The matched appearance vector is first projected into the low dimensional space and then to the RCA feature space by means of projection matrices learned in the training step. Expressions are classified in one of the 6+1 basic emotional categories. The supplementary class is added to take into account neutral expressions. The expression of the unseen face is assigned to the class of the nearest neighbor in the Euclidean distance sense chosen among the training examples.

The RCA-based expressions classifier is compared with some other linear and nonlinear methods.

In particular we compare our approach to the one proposed by Abboud et al. [1], in which the recognition is performed in the Fisherspace. They use linear discriminant analysis in order to extract the 6 most discriminating features which maximize class separability and compute the mean vector $\overline{\mathbf{c_i}}$ for each class. The tested face is assigned to the class having the nearest mean.

Concerning the nonlinear classifiers, we test our method against a nonlinear variation to the Abboud approach replacing the FLD with a generalized discriminant analysis(GDA)[4]. The GDA is a kernel-based method for nonlinear classification based on a mapping of the input space into a high dimensional feature space with linear properties. In the new space, one can solve the problem with the classical FLD method. We finally compare with a

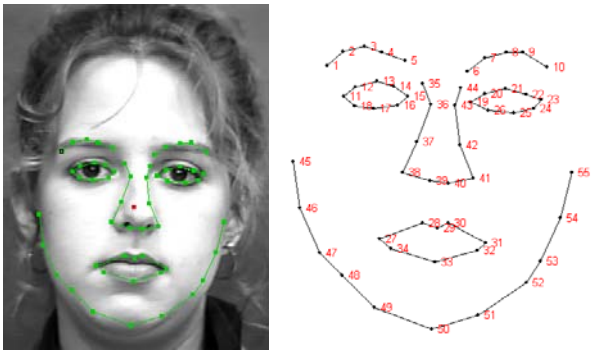

Figure 2. Facial landmarks (55 points)

well tuned c-SVM[21].

4. Experiments

In order to test the algorithms described above we use the Cohn-Kanade Database[13]. The database consists of expression sequences of subjects, starting from a neutral expression and ending most of the time in the peak of the facial expression. There are 104 subjects in the database, but only for few of them the six expressions are available. Our framework requires a training set to build the AAM model, a training set to learn the expressions manifold and a set of unseen faces to test its performance. The classifiers used in the comparison study will all share the same training and test sets. As mentioned in Section 3 the fundamental pre-processing step in the described framework consists in building an active appearance model. The appearance model is built using 300 images[1] from 11 different subjects chosen in the database. The AAM training set is composed by 48 neutral images and 42 images for each of the 6 primary emotions. The latest ones have been chosen considering emotions at different levels of magnitude. In order to build the model we have manually landmarked the images of the training set using the facial model showed in Fig.2. The model is built using 49 shapes model, 140 texture modes and 84 appearance modes, thus retaining the 98% of the combined shape and texture variation. The shape-free texture vector $\mathbf{g}$ is compose of 38310 pixels and the shape vector dimension is 55. Concerning the implementation, we use the C++ code of Active Appearance Model available on the AAM web page[2].
The classification training and test set consist respectively of 143 and 115 appearance masks. Table 4 shows the number of images for each expression in the expression training and test set.

[1]the list of the chosen subjects for training and test will be available on the web in case of acceptance

[2]http://www2.imm.dtu.dk/ aam/

Classifiers	Correct Classification Rate(%)
SVM	86.957
RCA	75.652
LDA	78.261
GDA	83.478
FLD+RCA	85.217
GDA+RCA	82.609

Table 1. Classification rate for 6 different classifiers

Expressions	Training images	Test images
Neutral	26	15
Happiness	20	18
Surprise	21	20
Fear	18	11
Anger	18	17
Disgust	22	17
Sadness	18	17

Table 4. Number of images in the classification training and test set

Table 1 shows the results on applying the different classifiers. We used the standard c-SVM implementation from `libsvm`[3]. We experimented with a range of polyomial, gaussian radial basis function (RBF) and sigmoid kernels and found that RBF kernels outperform the others. The tuning of the SVM has been performed initially by a cross-validation and afterwards by means of manual search. The RCA entry in Table 1 refers to the framework of Fig.1 omitting the dimensionality reduction step. The LDA classifier follows the framework described in [1], projecting features in Fisherspace of dimensionality 6. The GDA classifier, as mentioned in Section 3 is a kernel version of the previous one, keeping the dimension of the embedded feature space to 6. The kernel used is a third degree polynomial function. The good classification rate of our GDA version of the Abboud approach reveals a better representation of the facial manifold using this nonlinear technique. The last two lines of Table 1 show the results for the proposed approach, where FLD and GDA are considered as dimensionality reduction techniques. Analysing the values of the classification rates in Table 1, it turns out that the SVM classifier achieves the best recognition rate. However the results given by FLD+RCA are close to the best performing SVM. In contrast with the tedious and subjective tuning of the SVM, the FLD+RCA classifier is not affected by this time consuming step, while keeping a good recognition rate.
Another remarkable observation comes from the gap in the

[3]http://www.csie.ntu.edu.tw/ clin/libsvm

FLD+RCA	HAPPINESS	SURPRISE	FEAR	ANGER	DISGUST	SADNESS	NEUTRAL	*Overall(%)*
HAPPINESS	17	0	1	0	0	0	0	94.44
SURPRISE	0	19	0	0	1	0	0	95.00
FEAR	0	0	10	0	1	0	0	90.91
ANGER	1	0	0	7	3	1	5	41.18
DISGUST	1	0	0	0	16	0	0	94.12
SADNESS	0	0	0	1	0	16	0	94.12
NEUTRAL	0	0	1	0	0	1	13	86.67

Table 2. Confusion matrix for the FLD+RCA classifier

SVM	HAPPYNESS	SURPRISE	FEAR	ANGER	DISGUST	SADNESS	NEUTRAL	*Overall(%)*
HAPPINESS	18	0	0	0	0	0	0	100.00
SURPRISE	0	18	0	0	1	1	0	90.00
FEAR	0	0	10	0	1	0	0	90.91
ANGER	2	0	0	10	2	1	2	58.82
DISGUST	1	0	0	0	16	0	0	94.12
SADNESS	0	0	0	4	0	13	0	76.47
NEUTRAL	0	0	0	1	0	0	14	93.33

Table 3. Confusion matrix for the SVM classifier

recognition rate between FLD and FLD+RCA. This result is coherent with what Bar-Hillel et al.[2] obtained applying RCA to facial recognition. As in the face recognition application, the use of RCA dramatically enhances the performance of FLD.
Finally Tables 2 and 3 show the confusion matrices for the two best performing classifiers, FLD+RCA and SVM. We note that anger is the most confused expression. The explanation to this comes from the subtle appearance differentiation between anger and its corresponding misclassified expressions.

5. Conclusions

In this paper we presented and tested a facial expression recognition framework, using RCA as a mathematical tool to learn a good metric from the input data. The proposed system has been tested against some state of the art linear and nonlinear classification methods. Our second task, in this work, has been to benchmark some state of the art linear and nonlinear classifiers. Our results indicate that, though an ad-hoc well tuned SVM still gives slightly better recognition rate, the good performance and the "plug-&-play" nature of our approach make it a good trade-off between complexity and classification rate.
In the future work we will address the problem of the dynamic classification. The use of video sequences will certainly add a more discriminative power to the classification task. At the same time we will quantify the recognition performances of humans on the same tested video and images. The goal will be to study and compare the human and the machine misclassification distribution. A hybrid of classifiers using static and dynamic classification will also be part of our future research.

6. Acnowledgments

Our work is supported by the Swiss National Science Foundation through the National Centre for Competence in Research(NCCR) on Interactive Multimodal Information Management(IM2).

References

[1] D. Abboud and F. Davoine, "Appearance factorization based facial expression recognition and synthesis." in *ICPR (4)*, 2004, pp. 163–166.

[2] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a mahalanobis metric from equivalence constraints," *Journal of Machine Learning Research*, vol. 6, pp. 937–965, June 2005.

[3] M. Bartlett, "Face image analysis by unsupervised learning and redundancy reduction," 1998.

[4] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach." *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[5] D. Blum, "Faceit!" *Psychology Today*, pp. 32–66, October 1998.

[6] H. Changbo, Y. Chang, R. Feris, and M. Turk, "Manifold based analysis of facial expression," *Computer Vision and Pattern Recognition Workshop, 2004 Conference on*, vol. 27, pp. 81–?, June 2004.

[7] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. S. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," *Computer Vision and Image Understanding*, no. 10, pp. 160–187, July-August 2003.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, pp. 681–685, June 2001.

[9] C. Darwin, *The Expression of the Emotions in Man and Animals*. London: J.Murray, 1872.

[10] I. L. Dryden and K. V. Mardia, *Statistical Shape Analysis*. New York: Wiley, 1998.

[11] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, no. 17, pp. 124–129, 1971.

[12] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the Human Face*. Elmsdorf, NY: Pergamon Press, 1972.

[13] T. Kanade, J. Cohn, and Y. L. Tian, "Comprehensive database for facial expression analysis," in *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, March 2000, pp. 46 – 53.

[14] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic interpretation and coding of face images using flexible models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 743–756, 1997.

[15] J. Lien, "Automatic recognition of facial expressions using hidden markov models and estimation of expression intensity," 1998.

[16] K. Mase and A. Pentland, "Recognition of facial expression from optical flow," *IEICE Transactions*, no. E74, pp. 3474–3483, October 1991.

[17] P. Michel and R. E. Kaliouby, "Real time facial expression recognition in video using support vector machines," in *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces*. New York, NY, USA: ACM Press, 2003, pp. 258–264.

[18] C. Padgett and G. Cottrell, *Representing face images for emotion classification*. Cambridge, MA: MIT Press, 1997.

[19] N. Shental, T. Hertz, D. Weinshall, and M. Pavel, "Lecture notes in computer science," *Computer Vision and Pattern Recognition Workshop, 2004 Conference on*, 2002.

[20] M. B. Stegmann, "Active appearance models: Theory, extensions and cases," Master's thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, aug 2000.

[21] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer Verlag, 1995.