Upper Facial Action Unit Recognition

Cemre Zor and Terry Windeatt

Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey, Guildford, Surrey, United Kingdom GU2 7XH
{c.zor,t.windeatt}@surrey.ac.uk

Abstract. This paper concentrates on the comparisons of systems that are used for the recognition of expressions generated by six upper face action units (AUs) by using Facial Action Coding System ($FACS$). Haar wavelet, Haar-Like and Gabor wavelet coefficients are compared, using Adaboost for feature selection. The binary classification results by using Support Vector Machines (SVM) for the upper face AUs have been observed to be better than the current results in the literature, for example 96.5% for *AU2* and 97.6% for *AU5*. In multi-class classification case, the Error Correcting Output Coding ($ECOC$) has been applied. Although for a large number of classes, the results are not as accurate as the binary case, $ECOC$ has the advantage of solving all problems simultaneously; and for large numbers of training samples and small number of classes, error rates are improved.

Keywords: FACS, ECOC, Adaboost.

1 Introduction

Face expression analysis and recognition have been one of the fastest developing areas due to its wide range of real life application areas such as emotion analysis, biometrics, computer graphics / simulations and image retrieval. The task is challenging and requires research into solving problems occurring under different illuminations, orientations and numerous other variations. In this paper, the aim is to compare binary and multi-class face expression analysis algorithms using the Facial Action Coding System ($FACS$) [5], which allows division of the main problem into sub-problems containing specific Action Units (AUs). Features based on Haar-Like, Gabor and Haar wavelet coefficients are compared using feature selection method based on Adaboost. Classification is performed by Support Vector Machines (SVM) and the multi-class problem is solved with Error Correcting Output Coding ($ECOC$).

Recognition rates for binary classification have been presented on three benchmark datasets and for binary and multi-class on the Cohn-Kanade frontal comprehensive database for facial expression analysis [20].

In Section 1.1, brief background of the problem has been presented. Section 2 describes the normalization steps, and in Sections 3 and 4, the components of the implemented systems; feature extraction, selection and classification have been explained in detail. Finally in Section 5, binary and multi-class classification results can be examined.

M. Tistarelli and M.S. Nixon (Eds.): ICB 2009, LNCS 5558, pp. 239–248, 2009.

1.1 Background

In the literature, when facial expression analysis is considered, two main different approaches and two different parameterisations can be found. The first approach uses the whole frontal face image for classifying into six universal facial expression prototypes: disgust, fear, joy, surprise, sadness and anger. Ekman and Friesen have proposed in their related work that each of the six emotions has characteristic expression on the face and that's why recognition of them is necessary and sufficient [1][2]. Bartlett, Littlewort et al [3][4] have also used the method for fully automatic recognition systems.

As changes of features on specific sub-sections such as eyes and eyebrows are assumed to be more related to facial expression, dividing the face images into sub-sections for further processing is the main idea of the second approach. 'Facial Action Coding System', first developed by Ekman and Friesen [5]; for describing facial expressions by 44 different AUs widens the range of applications of face expression recognition and most current work on facial expression analysis depends on this decomposition into AUs.

There are also other methods that use neither the frontal face image as a whole nor all of the 44 AUs, but some other criteria such as the manually selected regions on face [6] or surface regions of facial features [7].

There are two main parameterisations that are used in both of the above approaches. Geometric based parameterization is an old way which consists of tracking and processing the motions of some spots on image sequences, firstly presented by Suwa [8] to recognize facial expressions. Cohn and Kanade later on tried geometrical modelling and tracking of facial features by claiming that each AU is presented with a specific set of facial muscles. In general, facial motion parameters [6][7] and the tracked spatial positioning & shapes of some special points [9] on face, are used as feature vectors for the geometric based method. These feature vectors are then used for classification. As for the disadvantages of this method; manual contour adjustments for the features on the initial image frames of each individual subject, lack of robustness occurring due to pose and illumination changes while the tracking is applied, and difficulties in the estimation of general movement / displacement parameters as a result of change in actions & expressions in morphological and dynamical senses [12], can be listed.

Rather than tracking spatial points and using positioning and movement parameters that vary within time, colour (pixel) information of related regions of the face are processed in appearance based parameterizations. Features such as Gabor, Haar wavelet coefficients, together with feature extraction and selection methods such as PCA, LDA and Adaboost are used within this framework. Example research can be found in [11][4].

The combination of the geometric and appearance based methods have also been used in some previous work. For example, Zhang [10] has tracked some fiducial points on the face images while also taking the Gabor wavelets of these points into account for the facial expression recognition.

In this paper, appearance based parameterization has been used together with $FACS$. Within the implementation of the binary classification scheme; Haar and

Gabor wavelet coefficients; and Haar-Like features have been used as extracted features. Feature selection has been applied by Adaboost, and *SVM* and Adaboost classifications have been performed in the final stage. When the multi-class classification problem is taken into consideration, the method used is the *ECOC* technique which is combined with the Adaboost feature selection and *SVM* classification techniques, together with an application of Bootstrapping on the training data.

2 Normalization

As the experiments were carried out on the upper face AUs, eye centres / pupils were localized on the input images as the first step of normalization. Localization was performed manually for the experiments so that the results were not dependent on the possible inaccuracies of an automated system. After the pupil coordinates were marked in each image, orientation and size normalizations were carried out through bilinear transforms. Images were made to have a distance of 32 pixels between the left and the right eye centres and a zero degree slope between the y-coordinates of these centres. Then, 32 by 32 square regions containing the eyes and the eyebrows were cropped around the pupils.

The previously detected, cropped, resized and rotated 32 by 32 gray scale images were then made to go through low pass filtering and a couple of histogram processing steps including contrast sketching and outlying pixels elimination. Elimination was accomplished by using the standard deviation information obtained from the histograms which look like scaled and shifted Gaussians. In addition to those, a log / power-law transformation was applied in order to have the average pixel value of each image equal to 128.

3 Feature Extraction and Selection

After the normalization step, Haar and Gabor wavelet coefficients, and Haar-Like features have been extracted in order to form up the feature vectors.

The superiority of Haar wavelet coefficients over the most commonly used Gabor wavelets and also the Haar-Like coefficients is their superiority in terms of extraction times. The mother wavelet function of Haar wavelets might be considered as a kind of step function:

$$\psi(x) = \left\{ \begin{array}{ll} 1, & 0 \leq t < 1/2 \\ -1, & 1/2 \leq t < 1 \\ 0, & otherwise \end{array} \right\}. \tag{1}$$

The wavelet decomposition of an image could therefore be defined as the combination of the resulting difference images calculated in different scales.

Haar-Like coefficients, introduced by Viola and Jones [13], are features which are indicatives of Haar basis functions. They consist of so-called two-rectangle, three-rectangle and four-rectangle features, which compose of differences between the sums of pixels of the same-size pairs of rectangles. Compared to Gabor

and Haar wavelets, it takes a longer time to extract all the Haar-Like features of an input image. For example, a 24*24 image has 160000 Haar-Like features while a 32*32 one has more than 450000. This means that the set of features is many times over-complete; however, any rectangular sum can be computed in only four array differences by making use of the "integral image" method that is expressed by Viola and Jones [13]. Haar-Like features are sensitive to edges, boundaries and other important information hidden in pixel values such as the difference between the pixel values on the regions of motion on face.

Gabor wavelet transformation has been used in various kinds of signal and pattern processing / analysis areas both in spatial and in frequency domains and is found to give satisfactory results in application areas such as texture segmentation [16], fingerprint recognition [17] and face recognition [15]. The characteristics of Gabor wavelets such as their ability to get easily adjusted for detailed localization in spatial and frequency domains [18] and the similarity between their frequency & orientation representations and the human visual system components have made them popular for particular usage areas. Gabor wavelets are formed from the multiplication of a complex sinusoidal carrier with a Gaussian envelope.

$$g(x,y) = K\exp(-\pi(a^2x^{'2} + b^2y^{'2}))\exp(j(2\pi F_o(x\cos w_o + y\sin w_o) + P)) \quad (2)$$

where

$$x^{'} = (x - x_0)\cos\theta + (y - y_0)\sin\theta;\;\; y^{'} = (y - y_0)\cos\theta - (x - x_0)\sin\theta$$

and where K is the scale parameter for the magnitude of the envelope, a and b the parameters of the x and y axis of the envelope, θ the envelope rotation angle, F_o and w_o the magnitude and the direction of the sinusoidal carrier's frequency, P the phase of the carrier, and x_o and y_o the location of the peak value of the envelope.

The most commonly used values of the phase in the literature for facial expression recognition and also face recognition have been used here too: $P = \pi\frac{i}{8}; i = 0, ..., 7$. In case of frequencies, the upper limit is equal to 0.5 Hertz due the Nyquist sampling theory; and because having $\sqrt{2}$ as the spacing between the kernels in the frequency domain is found to give out experimentally good results on previous research, the values of F used are equal to: $\frac{fmax}{\sqrt{2}^u}, u = 0, ..., 4$ where $fmax = 0.5$.

For feature selection, the Adaboost algorithm has been used. Adaboost was originally a strong and fast classification algorithm, introduced by Freud and Schapire [14], making use of a weak binary classifier that strengthens its decisions in each iteration to end up with a final hypothesis with the lowest error rate. The weak binary classifier (weak learner) is any classifier for which the weighted classification error is expected to be better than chance. More recently, particularly in the Computer Vision community, Boosting has become popular as a feature selection routine, in which a single feature is selected in each Boosting iteration [13]. Specifically, the Boosting algorithm is modified so that, in each iteration, the individual feature which minimises the classification error on the

weighted samples is chosen [13]. In our implementation, we use Adaboost with decision stump as weak learner.

4 Classification

Support Vector Machines (*SVM*), which were firstly developed from Statistical Learning Theory by Boser, Guyon and Vapnik [23], are used as binary classifiers in this paper. They aim to perform 2-class classification via optimally separating the data by making use of an N-dimensional hyper-plane. While dealing with the optimal hyper-planes that are to classify the data by minimizing the empirical classification error, *SVM* take also the maximization of the margin, in other words achievement of the maximum separation into account.

For multi-class classification, Error Correcting Output Coding (*ECOC*) has been used. In the *ECOC* matrix, every row represents the unique codeword (a binary string of n) of a class. For each of the n columns, a binary classifier, which is *SVM* here, has been trained. The features extracted are Gabor wavelets and the feature selection method is Adaboost. Also, in order to reduce correlation between individual columns, bootstrapping [21] is applied during training.

Each of the n trained *SVM* are applied to the new test pattern, and the resulting binary values are combined to create its binary codeword. This codeword is then compared to each of the r codewords and classified according to the codeword with closest distance, using Hamming distance [19].

One of the advantages of *ECOC* is the method's ability to correct bit errors up to a degree. If the minimum Hamming distance between any pair of codewords is d, then at least $\lfloor (d-1)/2 \rfloor$single bit errors can be corrected. Therefore, row separation in the design of the error correcting output code matrix is quite important for the classification results. Also, in order to create a good *ECOC* matrix, the Hamming distance between a pair of columns should be large enough so that the deterministic learning functions per each column are uncorrelated [19]. In our approach, the additional usage of bootstrapping helps further to de-correlate the columns.

5 Results

5.1 Binary Classification

Experiments were carried out on the Cohn-Kanade frontal face database [20] together with additional supplementary datasets from UCI Machine Learning Repository [22]. The images in the Cohn-Kanade dataset have been coded using *FACS*, which describes subject's expression in terms of *AU*s [20].

In the first part of the experiments, comparison of the two Adaboost algorithms, Adaboost as a classifier without feature selection (*Ada*); and with feature selection (*AdaFs*) have been carried out on small datasets, Breast Cancer Wisconsin (*Breast-w*), Sonar Mines vs. Rocks (*Sonar*), and Pima Indians Diabetes (*Diabetes*) which were taken from the UCI Machine Learning Repository. The

Breast-w dataset consists of 699 data with 10 features whereas the *Sonar* dataset consists of 208 data with 10 features, and the *Diabetes* of 768 data with 8 features. The number of features chosen by *AdaFs* was half of the number of total features within each dataset. In case of Diabetes and Sonar datasets, error rates were lower when *AdaFs* was used instead of *Ada*, whereas both of the rates for Breast-w dataset were similar. Due to these rates and also its ability to carry out feature selection at the same time with classification, *AdaFs* was considered to be superior to *Ada*.

The correct classification rates of *AdaFs* have therefore been compared to *AdaFs + SVM* combination in the second step. In *AdaFs + SVM*, half of the features were firstly selected from the data by *AdaFs* and the classification itself was carried out by *SVM*. *AdaFs + SVM* was better in terms of recognition rates, and Table 1 lists the error percentage results for the comparison of the three classification methods using 10-fold cross validation.

Table 1. Recognition error rates (%) of *Ada*, *AdaFs* and *AdaFs+SVM*, applied on UCI MLR

	Ada	*AdaFs*	*AdaFs + SVM*
Breast-w	4.0	4.6	3.5
Diabetes	28.1	26.7	22.5
Sonar	19.6	14.7	11.0

For the Cohn Kanade dataset *AdaFs + SVM* was found to give the best results, having an average of 5% increase over the performance of *Ada*, and 2% over *AdaFs*. Here, Haar wavelets were used as feature vectors and the number of features selected was set to 100 as there was a decrease in performance of about 1.5% when 200 and 300 were used. Also, there is the advantage of a decrease in training and run times for 100 features.

Though revealing good results, Haar wavelet coefficients depend on the differences between some determined pixel intensities; therefore even 1-2 pixel shifts that occur due to non-robust localizations of the eye centers may affect the overall results. Therefore Haar-Like features and Gabor wavelets were considered. Haar-Like features implement the idea of using the change in the contrast values between adjacent rectangular groups of pixels instead of the intensity values of individual pixels. Gabor wavelets are more robust to pixel intensity and lighting changes than Haar wavelets and are also meaningful in frequency domain. In Table 2, recognition rates for *AdaFs+SVM* (number of selected features being equal to 100) have been given for Haar wavelet, Haar-Like and Gabor wavelet coefficients on the upper face *AU*s of the 32 by 32 normalized Cohn-Kanade dataset images.

Although the resulting classification rates of Haar-Like features are not as successful compared to Haar wavelets, in real life applications where the localization of the centres are done automatically and where inaccurate localizations are more likely to occur, Haar wavelet features are expected to reveal poorer performance.

Table 2. Recognition error rates (%) obtained by Gabor, Haar wavelet coefficients, and Haar-Like features

	Haar Wavelet Coefficients	Haar-Like Features	Gabor Wavelet Coefficients
AU1	9.0	17.0	5.2
AU2	4.5	10.5	3.3
AU4	12.0	22.0	8.1
AU5	3.1	8.3	2.4
AU6	11.6	19.5	8.3
AU7	12.9	19.1	8.5

In case of Gabor wavelets, it was observed for all of the AUs that the first 10 out of 100 features used for classification have been extracted by using Gabor filters with mostly low complex sinusoidal frequencies. This is therefore an indication of the fact that useful frequency bands for facial AUs recognition are the low frequency bands in images. However, future work is needed to reveal why Gabor features do not make much use of edge information. One possible explanation might be that although convolution is taken into account, edge information might vary quite a lot in images even if they belong to the same class; and therefore this information may not be useful, and maybe even redundant. It was also found that both the number of features used and the parameters selected for Gabor wavelets play important roles in the final classification results. 100 features were found to give the optimal results together with the detected wavelet parameters: a=1/ 50, b=1/ 40 and the width and the height of the filter size being equal to half of the width and height of the image to be filtered. Experiments were also tested on 16 by 16 down-sampled images, and also on 1/10th down-sampled feature space; and the classification rates were found to be on average 5% less successful.

Although the classification rates for Gabor wavelets came out to be the most successful, Whitehill and Omlin [11] showed that extraction of Gabor wavelet coefficients is 300 times more costly than Haar. However, the down-sampled images or features might be used in real-time applications where speed is of importance.

5.2 Multi-class Classification

In order to carry out multi-class classification of the AUs, $ECOC$ is used. As Gabor features with $AdaFs + SVM$ classification on 32 by 32 images were found to give the best binary classification, these components were inherited in the $ECOC$ system. The strategy consists of applying bootstrapping, as explained in Section 5, followed by $AdaFs + SVM$ on Gabor features for each column. Tables 3, 4 and 5 show the results of experiments with ten-fold cross validation that were applied to three different sets of data, one having 12, the other 4 and the last one having 3 classes. In the set of experiments, in order to decrease the time needed for training, every 1 out of 10 Gabor features from a total of 40960 was taken to form the new feature set, to which $AdaFs$ was applied to select 100 features.

Table 3. Recognition error and true positive rates (%) obtained by *ECOC* with 12 classes

Class No	AUs that are "on"	Number of Data	Overall Error / Tp Rate
1	1	7	5.3 / 0.0
2	4	26	12.1 / 17.8
3	6	65	14.6 / 62.9
4	7	5	3.0 / 0.0
5	1 & 2	23	8.4 / 20.0
6	1 & 4	20	6.6 / 42.0
7	4 & 7	47	16.0 / 61.1
8	6 & 7	13	6.5 /14.2
9	1 & 4 & 7	11	6.1 / 4.7
10	1 & 2 & 4	6	4.0 / 0.0
11	1 & 2 & 5	62	7.0 / 76.0
12	4 & 6 & 7	22	10.4 / 4.7

Table 4. Recognition error and true positive rates (%) obtained by *ECOC* with 4 classes

Class No	AUs	No. of Data	12 Class Err. / Tp	4 Class Err. / Tp
2	4	26	12.1 / 17.8	22.4 / 33.6
3	6	65	14.6 / 62.9	13.6 / 69.7
7	4 & 7	47	16.0 / 61.1	18.5 / 60.4
11	1 & 2 & 5	62	7.0 / 76.0	4.2 / 97.6

In Table 3, all the possible combinations of the upper face AUs that exist on the Cohn-Kanade dataset images were used to form 12 classes. The classes were created to be mutually exclusive and the ones having less than 5 data samples were discarded. Dealing with a 12 class problem is a complicated and challenging task and thus, in this stage we have ended up with an average 25.2% true positive (*tp)* rate, although the average error rate is 8.3%.

Apart from the number of classes, the effect of the number of training patterns on classification should also be considered. In the 12 class problem, the best *tp* rates were obtained by Class 11 and Class 3, and these classes are ones with the maximum number of patterns. Also, the worst ones were obtained by Class 1, Class 4 and Class 10, which have only 7, 5 and 6 patterns respectively. From Table 4, the four classes containing the maximum number of patterns, Class 11, Class 3, Class 7 and Class 2, can be seen to give better *tp* rates than the 12 class case. However, main reasons for some misclassifications are believed to be due to the low number of patterns in Class 2, and to the fact that two of the classes, Class 2 and Class 7 contain the same AU: $AU7$, which might cause confusion. Therefore the experiments shown in Table 5 were applied to the 3 class problem, which excludes Class 2. Since the classes now all have exclusive AUs, with only classes having the maximum number of patterns included, recognition and the *tp* rates are much improved on average.

Table 5. Recognition error and true positive rates (%) obtained by *ECOC* with 3 classes

Class No	*AU*s	No. of Data	12 Cl. Err. / Tp	4 Cl. Err. / Tp	3 Cl. Err./ Tp
3	6	65	14.6 / 62.9	13.6 / 69.7	12.9 / 76.5
7	4 & 7	47	16.0 / 61.1	18.5 / 60.4	6.4 / 95.4
11	1 & 2 & 5	62	7.0 / 76.0	4.2 / 97.5	8.0 / 97.5

6 Conclusion

In this paper both the binary and the multi-class classification results of the upper face action units have been presented through making use of several systems. In the binary case, the overall results were observed to be better than current results in the literature.

The system that was found to give out the best binary classification results was used in the multi-class case with *ECOC*. Although the recognition rates using twelve classes were low as expected due to classes have common *AU*s and insufficient training data, the results for smaller number of classes with large number of training data were much improved. Therefore the results are highly dependent on how the problem is decomposed and particularly if the classes are mutually exclusive. Future work is aimed at finding the best combination of feature selection and *ECOC* coding and decoding strategies for facial action unit recognition.

References

1. Ekman, P., Friesen, W.V.: Pictures of Facial Affect. Consulting Psychologist Press, Palo Alto (1976)
2. Izard, C., Dougherty, L., Hembree, E.A.: A System for Identifying Affect Expressions by Holistic Judgements. Univ. Of Delaware (unpublished manuscript) (1983)
3. Bartlett, M.S., Hager, J., Ekman, P., Sejnowski, T.: Measuring Facial Expressions by Computer Image Analysis. J. Psychophysiology 36, 253–263 (1999)
4. Bartlett, M.S., Littlewort, G., Lainscsek, C., Fasel, I., Movellan, J.: Machine Learning Methods for Fully Automatic Recognition of Facial Expressions and Facial Actions. In: IEEE International Conference on Systems, Men and Cybernetics, Netherlands, pp. 592–597 (2004)
5. Ekman, P., Friesen, W.V.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologist Press, Palo Alto (1978)
6. Mase, K.: Recognition of Facial Expression from Optical Flow. IEICE Trans. E74(10), 3474–3483 (1991)
7. Yacoob, Y., Davis, L.S.: Recognizing Human Facial Expression from Long Image Sequences Using Optical Flow. IEEE Trans. Pattern Analysis and Machine Intelligence 18(6), 636–642 (1996)
8. Suwa, M., Sugie, N., Fujimora, K.A.: Preliminary Note on Pattern Recognition of Human Emotional Expression. In: Proc. International Joint Conf. Pattern Recognition, pp. 408–410 (1978)

9. Lanitis, A., Taylor, C., Cootes, T.: Automatic Interpretation and Coding of Face Images Using Flexible Models. IEEE Trans. Pattern Analysis and Machine Intelligence 19(7), 743–756 (1997)
10. Zhang, Z.: Feature-Based Facial Expression Recognition: Sensitivity Analysis and Experiments with a Multilayer Perceptron. Int'l. J. Pattern Recognition and Artificial Intelligence 13(6), 893–911 (1999)
11. Whitehill, J., Omlin, C.W.: Haar Features for FACS AU Recognition. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (2006)
12. Donato, G., Bartlett, M.S., Hager, J., Ekman, P., Sejnowski, T.J.: Classifying Facial Actions. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(10), 974–988 (1999)
13. Viola, P., Jones, M.J.: Robust Real-Time Face Detection. International J. of Computer Vision 57(2), 137–154 (2004)
14. Freund, Y., Schapire, R.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Computer and System Sciences 55, 119–139 (1997)
15. Shen, L., Bai, L., Fairhurst, M.: Gabor Wavelets and General Discriminant Analysis for Face Identification and Verification. Image Vision Computing 25(5), 553–563 (2007)
16. Jain, A.K., Farrokhnia, F.: Unsupervised Texture Segmentation Using Gabor Filters. Pattern Recognition 24(12), 1167–1186 (1991)
17. Lee, C.J., Wang, S.D.: Fingerprint Feature Extraction Using Gabor Filters. Electronics Letters 35(4), 288–290 (1999)
18. Zhan, Y., Niu, D., Cao, P.: Facial Expression Recognition Based on Gabor Wavelet Transformation and Elastic Templates Matching. In: Third International Conference on Image and Graphics (ICIG 2004), pp. 254–257 (2004)
19. Dietterich, T.G., Bakiri, G.: Solving Multi-class Learning Problems via Error-Correcting Output Codes. J. Artificial Intelligence Research 2, 263–286 (1995)
20. Tian, Y., Kanade, T., Cohn, J.F.: Recognizing Action Units for Facial Expression Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(2), 97–115 (2001)
21. Efron, B.: Bootstrap methods: Another Look at the Jackknife. The Annals of Statistics 7(1), 1–26 (1979)
22. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. School of Information and Computer Science, University of California, Irvine (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
23. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A Training Algorithm for Optimal Margin Classifiers. In: COLT 1992: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144–152. ACM Press, New York (1992)