Automatic Facial Expression Recognition Using Boosted Discriminatory Classifiers

Stephen Moore and Richard Bowden

Centre for Vision Speech and Signal Processing
University of Surrey, Guildford, GU2 7JW, UK
{stephen.moore,r.bowden}@surrey.ac.uk

Abstract. Over the last two decades automatic facial expression recognition has become an active research area. Facial expressions are an important channel of non-verbal communication, and can provide cues to emotions and intentions. This paper introduces a novel method for facial expression recognition, by assembling contour fragments as discriminatory classifiers and boosting them to form a strong accurate classifier. Detection is fast as features are evaluated using an efficient lookup to a chamfer image, which weights the response of the feature. An Ensemble classification technique is presented using a voting scheme based on classifiers responses. The results of this research are a 6-class classifier (6 basic expressions of anger, joy, sadness, surprise, disgust and fear) which demonstrate competitive results achieving rates as high as 96% for some expressions. As classifiers are extremely fast to compute the approach operates at well above frame rate. We also demonstrate how a dedicated classifier can be consrtucted to give optimal automatic parameter selection of the detector, allowing real time operation on unconstrained video.

1 Introduction

Our objective is to detect facial expressions in static images. This is a difficult task due to the natural variation in appearance between individuals such as ethnicity, age, facial hair and occlusion (glasses and makeup) and the effects of pose, lighting and other environmental factors. Our approach relies upon a boosted discriminatory classifier based upon contour information. Contours are largely invariant to lighting and, as will be shown, provide an efficient discriminatory classifier using chamfer matching.

Given a cartoon or line drawing of a face, it is taken for granted that the human brain can recognize the expression or emotional state of the character. Sufficient information must therefore be present in this simplified representation for a computer to recognize key features associated with expressions. Using only contour information provides important advantages as it offers some invariance to lighting and reduces the complexity of the problem. Our approach relies upon finding edges/contours on the face that are consistent across individuals for specific expressions. From a large set of facial images, candidate edges are extracted

S.K. Zhou et al. (Eds.): AMFG 2007, LNCS 4778, pp. 71–83, 2007.

and a subset of consistent features selected using boosting. The final classifier is then a set of weighted edge features which are matched to an image quickly using chamfer-matching resulting in a binary classifier that detects specific emotions.

This paper is split into a number of sections, firstly a brief background of research into automatic facial expression recognition is presented. Section 3 explains the methodology of this research. Section 4 evaluates different expression classifiers and results. Real time implementation of this work is described in section 5 where the work is adapted for robust, continuous use. Parameterisation is addressed through the use of a dedicated classifier for automatic selection. Finally conclusions and future work are described in section 6.

2 Background

Automatic facial expression research has gained inertia over the last 20 years. Furthermore, recent advances in the area of face recognition and tracking, coupled with relatively inexpensive computational power has fueled recent endeavors.

Early work on Automatic Facial expression recognition by Ekman [8], introduced the Facial Action Coding System (FACS). FACS provided a prototype of the basic human expressions and allowed researchers to study facial expression based on an anatomical analysis of facial movement. A movement of one or more muscles in the face is called an action unit (AU) and all expressions can then be described by a combination of one or more of 46 AU's.

Feature extraction methods applied to facial expression recognition can be categorized into two groups, deformation methods or motion extraction methods. Deformation methods applied to facial expression recognition include Gabor wavelets [3] [6] [21], neural networks (intensity profiles) [1] and Active Appearance Models [15]. Gabor wavelets have achieved very accurate results as they are largely invariant to lighting changes and have been widely adopted in both facial detection and recognition, but are computationally expensive to convolve with an image. Motion extraction methods using optical flow [20] or difference images [5] have also been applied to facial expression recognition. Essa and Pentland [9] combined these approaches and demonstrated accurate recognition using optic flow with deformable models. This work also introduced FACS+, an extension of FACS into the temporal domain.

Expression recognition is closely related to face detection, and many approaches from detection (such as the Gabor methods previously mentioned) have been applied to expression recognition. Since the popularization of boosting in the vision community by Viola and Jones [17], this type of machine learning has received considerable attention. In Adaboost, a strong classifier is built as a simple linear combination of seemingly very weak classifiers. Viola and Jones built a fast and reliable face detector using Adaboost from simple, weak classifiers based upon 'haar wavelet like' block differences [17]. It is arguably the current state-of-the-art in face detection and has resulted in boosting being applied to many computer vision problems with many variants to the learning algorithm [13], [19]. Wang et al [18] extended this technique to facial expression recognition

by building separate classifiers of 'haar like' features for each expression. Shan and Gong [4] also applied boosting to facial expression recognition, but boosted local binary patterns (LBP) using conditional mutual information based boosting (CMIB). CMIB learns a sequence of weak classifiers that maximize their mutual information.

Shotton and Blake [16] presented a categorical object detection scheme based upon boosted local contour fragments. They demonstrate that the boundary contour could be used efficiently for object detection. This paper shows how internal contour features can be used for extremely fast discriminatory classification.

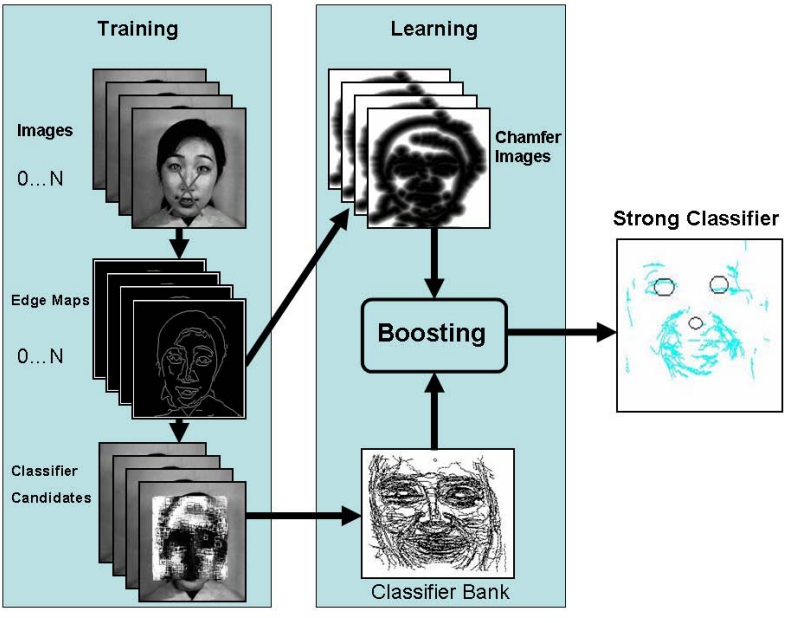

Fig. 1. System Overview

3 Methodology

3.1 Overview

In this section we introduce how the proposed approach works, illustrated in figure 1. A training set of images is extracted from a FACS encoded database. Images are annotated (eyes and tip of the nose) so that features can be transformed to a reference co-ordinate system. Each image then undergoes edge detection. From each edge image, small coherent edge fragments are extracted from the area in and around the face. A classifier bank (figure 2) is then assembled from candidate edge fragments from all the training examples. A weak classifier is formed by assembling an edge fragment combined with a chamfer score.

Boosting is then used to choose an optimal subset of features from the classifier bank to form a strong discriminatory classifier. The final boosted classifier provides a binary decision for object recognition. To build an n-class discriminatory classifier we use probability distributions built from classifier responses to allow likelihoods ratios to be used to compare across different classifiers. Also an investigation of fusion methodologies, a one against-many classifier and an ensemble classifier [7] are presented.

3.2 Image Alignment

To overcome the problem of image registration, each facial image must be transformed into the same co-ordinate frame. Our initial tests are performed using a 3 point basis. However we will then proceed to demonstrate that position and scale of a face (obtained via detection) is sufficient for classification with a minimal loss of accuracy. Before training, the images are manually annotated to identify the two eyes and the tip of the nose, to form a 3-point basis (points are non-collinear). Only near frontal faces are considered in this work and therefore a 3-point basis is sufficient to align examples.

3.3 Weak Classifiers

Expressions are based on the movement of the muscles, but visually we distinguish expressions by how these features of the face deform.

The contour fragments $e \in E$, where E is the set of all edges, are considered from the area around the face based on heuristics of the golden ratio of the face. The distance between the eyes is approximately half the width of the face and one third of the height. This identifies the region of interest (ROI) from which contours will be considered. Following an edge detection, connected component analysis is performed and from each resulting contour fragment, the contour is sampled randomly to form short connected edge features. Figure 2 shows an example of a classifier bank built from a training set of faces.

Fig. 2. Classifier Bank

3.4 Chamfer Image

To measure support for any single edge feature over a training set we need some way of measuring the edge strength along that feature in the image. This can be computed efficiently using Chamfer matching. Chamfer matching was first introduced by Barrow et al [2]. It is a registration technique whereby a drawing consisting of a binary set of features (contour segments) is matched to an image. Chamfer matching allows features to be considered as point sets and matching is efficient as the image is transformed into a chamfer image (distance) only once and the distance of any feature can then be calculated to the nearest edge as a simple lookup to that chamfer image. The similarity between two shapes can be measured using their chamfer distance.

All images in the training set undergo edge detection with the canny edge detector to produce an edge map. Then a chamfer image is produced using a distance transform DT. Each pixel value q, is proportional to the distance to its nearest edge point in E:

$$DT_E(q) = min_{e \in E} \|q - e\|_2 \tag{1}$$

To perform chamfer matching, two sets of edges are compared. The contour fragment (T) and image edge strength E, producing an average Chamfer score:

$$d_{cham}^{(T,E)}(x) = \frac{1}{N} \sum_{t \in T} min_{e \in E} \|(t + x) - e\|_2 \tag{2}$$

where N is the number of edge points in T. This gives the Chamfer score as a mean distance between feature T and the edge map E. Chamfer images are expensive to compute, however this needs only be computed once per image. The function $d_{cham}^{(T,E)}(x)$ is an efficient lookup to the chamfer image for all classifiers. An example of a chamfer image is shown in figure 1.

3.5 Learning

Boosting is a machine learning algorithm for supervised learning. Boosting produces a very accurate (strong) classifier, by combining weak classifiers in linear combination. Adaboost (adaptive boosting) was introduced by Freund and Schapire [10] and has been successfully used in many problems such as face detection [17] and object detection [16]. Adaboost can be described as a greedy feature selection process where a distribution of weights are maintained and associated with training examples. At each iteration, a weak classifier which minimizes the weighted error rate is selected, and the distribution is updated to increase the weights of the misclassified samples and reduce the weights of correctly classified examples. The Adaboost algorithm tries to separate training examples by selecting the best weak feature $h_j(x)$ that distinguish between the positive and negative training examples.

$$h_j(x) = \left\{ \begin{array}{l} 1 \; if \; d_{cham}^{(T,E)}(x) < \theta_j \\ 0 \; otherwise \end{array} \right\} \tag{3}$$

θ is the weak classifier threshold. Since setting a fixed threshold requires a priori knowledge of the feature space, an optimal θj is found through an exhaustive search for each weak classifier. An image can have up to 1,000 features, thus over the training set, many thousands of features are evaluated during the learning algorithm. This allows the learning algorithm to select a set of weak classifiers with low thresholds that are extremely precise allowing little deviation, and weak classifiers with high thresholds which allows consistent deformation of the facial features. This increases the performance but as will be seen, does not result in over fitting the data.

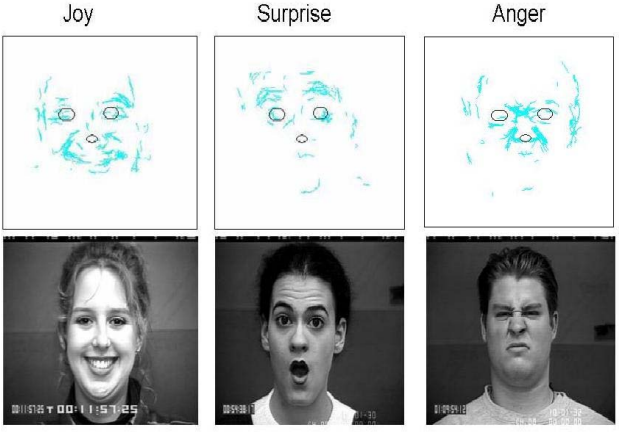

Fig. 3. Strong Classifier Visualization

Positive training examples are taken from the target expression and negative examples from other sets of expressions. Following boosting the final strong classifier consists of edge features which can be visualized. Figure 3 shows the classifiers for joy, surprise and anger trained against neutral expressions, the circles depict the position of the 3 point basis. Note that these visualizations reflect what we assume about expressions, eg surprise involves the raising of the eyebrows and anger 'the deformation' around the nose. However perhaps surprisingly, the mouth does not play an important role in the joy classifier, which is both counter intuitive and contradictory to AU approaches. This is partly due to higher variability away from the center of the 3 point basis, but more importantly the variability across subjects. People smile with their mouth open or closed, so boosting decides that the lines on the cheeks are more consistent features than those of the mouth. What boosting is doing is deciding its own optimal set of AU's based upon the data.

Training expressions against only neutral images results in a classifier that learns all the deformation for that expression. While this is beneficial for visualisation or single class detection it presents problems to multi class detection as many positive expressions will be confused by the classifiers. Training against all

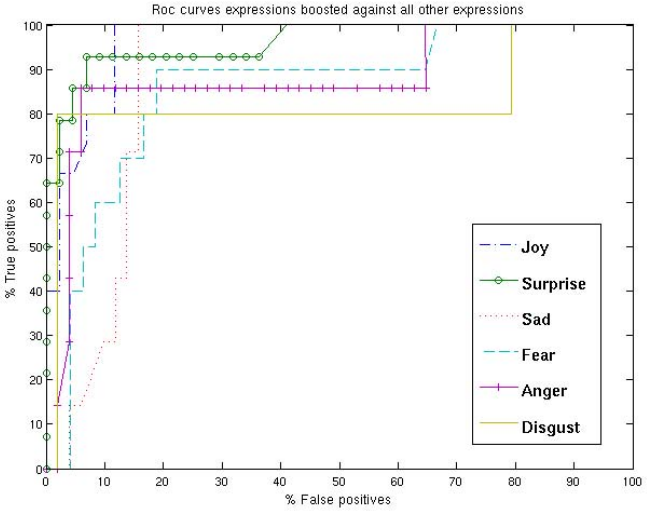

Fig. 4. Roc curves for each expression trained one against many

other expressions forms classifiers that only learn the deformation that is unique to that expression, which reduces the number of false positives. Figure 4 shows receiver operating characteristic (ROC) curves for each of the expression classifiers. Expressions were boosted using all other expressions as negative examples and over 1000 rounds of boosting.

3.6 N-Class Discriminatory Classifier

The following section is an investigation into different classifier approaches. In this paper we investigated using the threshold response from the strong classifier, likelihoods and ensemble methods for classification.

As our n-class classifier is to be built from binary classifiers some way of combining classifiers responses is required in order to disambiguate between expressions. The unthresholded classifier response cannot be used, as each classifier has a different number of weak classifiers, different thresholds and therefor different responses and ranges.

A more principled way to compare responses is to use likelihoods. Positive and negative probability distribution functions (pdf's) were constructed for each classifier, using a validation set. Noise in terms of x,y translation was added to the validation set in order to artificially increase the set. Positive and negative responses from the validation set were then used to build histograms (figure 5 and figure 6). Parzen windowing was used to populate these histograms. To calculate the likelihoods a comparison is made between the response of the positive pdf's for each classifier.

The likelihood ratio was evaluated for each classifier by dividing the response of the positive pdf by the response of the negative pdf for each classifier (equation

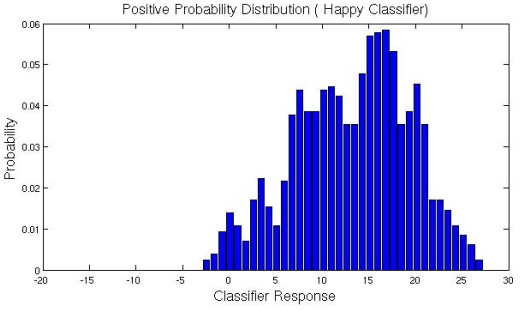

Fig. 5. Positive Probability Distribution

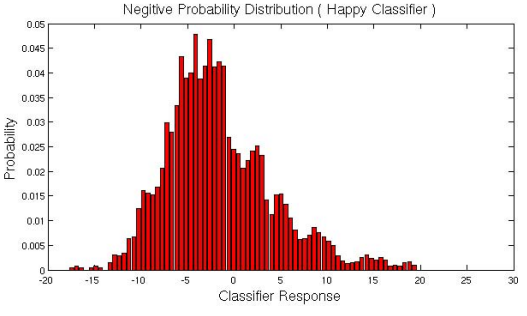

Fig. 6. Negative Probability Distribution

[4]). Where LR is the likelihood ratio, L is the likelihood and the positive and negative pdf's are Pp and Pn respectively.

$$LR(x) = MAX_{\forall n} \left\{ \frac{L(x, Pp)}{L(x, Pn)} \right\} \tag{4}$$

Dietterich [7] argues that ensembles methods can often perform better than a single classifier. [7] proposes three reasons why classifier ensembles can be beneficial (statistical, computational and representational). Statistical reasons are based on the learning algorithm choosing a non-optimal hypothesis, given insufficient training data. By constructing an ensemble from accurate classifiers, the algorithm can average the vote and reduce the risk of misclassification. For a n-class classifier system, this can be broken into $\frac{n(n-1)}{2}$ binary classifiers respectively, allowing each expression to be exclusively boosted against every other expression. Using a binary threshold each classifier has a vote. Each n expression ensemble classifier can receive (n-1) votes, and classification is done using a winner takes all strategy.

4 Expression Classification

The Cohn Kanade facial expression database [12] was used in the following experiments. Subjects consisted of 100 university students ranging in age from 18 - 30. 65% were female, 15% were African American, and three percent were Asian or Latino. The camera was located directly in front of the subject. The expressions are captured as 640 x 480 eps images. In total 365 images were chosen from the database. The only criteria was that the image represented one of the prototypical expressions. This database is FACS encoded and provides ground truth for experiments. Each image has a FACS code and from this code images are grouped into different expression categories. This dataset and the selection of data was used to provide a comparision between other similiar expression classifiers [4] and [3].

Initial experiments were carried out by training each expression against 1) neutral expressions only, 2) against all other expressions, selecting candidate features from positives training examples only and 3) against all other expressions, selecting negative and positive features from all images in the training set. Training expressions against only neutral images results in a classifier with the poorest performance as there is little variance in the negative examples and many other expressions are misclassified by the detector. Training against all other expressions improves performance as the classifier learns what deformation is unique to that expression. The better classifier is one that selects negative features to reduce false detections. This classifier outperforms the other two methods as each expression has unique distinguishing features which act as negative features. To give a crude baseline we normalize the classifier responses into the range 0-1 and the highest response wins. As expected likelihoods is a better solution with marginal performance gains. However the Likelihood ratio gives a significant boost . Using 5-fold cross validation on the 6-basis expressions and 7-class (neutral class included) a recognition rate of 67.69% and 57.46% is achieved.

Table 1. Recognition results 6 class

Method	*Joy*	*Surprise*	*Sad*	*Fear*	*Anger*	*Disgust*	*Overall*
Classifier Response	78.67	81.43	55.72	38	77.14	40	61.83
Likelihood	78.67	82.86	57.14	56	60	40	62.45
Likelihood Ratio	90.67	91.43	51.43	36	88.57	48	67.69
Ensemble Classifier Response	96	95.72	82.86	72	91.43	72	85

The recognition results were poor when compared to the roc curves (figure 4) for the classifiers. This is because when confusion between classifiers occurs, examples are misclassified. To overcome this confusion several more principled approaches were evaluated. Table 1 and table 2 show results using likelihoods and likelihood ratio's. All results presented in table 1 and table 2 are obtained using 5-fold cross validation with training and test sets divided 80-20. As expected,

likelihood ratios outperform likelihoods yielding a 5% increase in performance. From the results it was apparent that the more subtle expressions (disgust, fear and sad) are outperformed by expressions with a large amount of deformation (Joy, surprise,anger). Subtle changes in appearance are difficult to distinguish when using one reference co-ordinate frame due to the variability across subjects.

Table 2. Recognition results 7 class

Method	*Joy*	*Surprise*	*Sad*	*Fear*	*Anger*	*Disgust*	*Neutral*	*Overall*
Classifier Response	78.67	68.57	48.57	50	65.71	12	46	52.78
Likelihood	70.69	71.43	25.71	44	71.43	20	64	52.47
Likelihood Ratio	73.35	68.57	31.43	50	82.85	32	64	57.46
Ensemble Classifier Res	95.99	92.86	65.71	58	92.28	84	76	80.69

In this research we have a 6-class and 7-class classifier system, this can be broken down into 15 and 21 binary classes respectively, allowing each expression to be exclusively boosted against every other expression. Using a binary threshold (chosen from the equal error rate on the ROC curve) each classifier has a vote. Each n expression ensemble classifier can receive (n-1) votes. When confusion occurs, a table of likelihood responses is kept, the product of these is compared for each class of confusion and the highest likelihood wins. Using the binary voting scheme with the ensemble classifier gives an increase of up to 27% in recognition performance.

Table 3 compares this work with other facial expression classifiers. For a direct comparison we compare our results to other methods that use Adaboost and the Cohn Kanade database. Bartlett et al [3] performed similar experiments on the same dataset using Adaboost to learn Gabor wavelet features and achieved 85% accuracy. Further more Shan and Gong [4] learnt LBP through Adaboost and achieved 84.6% accuracy. Table 3 summarises that using contour fragments as a discriminatory classifier is comparably to Gabor wavlet and LBP features. It is important to note that while performance equals the state of the art, the application of our classifier is extremely efficient. A worst case classifier of 1000 weak classifiers takes only 3ms to assess within a half pal image based upon our implementation on a 3GHz P4 machine.

Table 3. Comparisons between other boosting based expression classifiers using Cohn Kanade database

Methods	*Results*
Local Binary Patterns with Adaboost [4]	84.6%
Gabor wavelets with AdaBoost [3]	85%
Edge/chamfer features with Adaboost	85%

5 Real Time Implementation

For real time detection in video sequences this work has been implemented with the Viola-Jones [17] face detector from the openCV libraries [11]. The initial experiments above required annotating the 3 point basis. For this implementation we use a 2 point basis from the bounding box returning by the face detector. Figure 7 shows the comparison of the three point basis (two eyes and nose) and the two point basis (points returning by face detector). Interestingly only a small performance drop is seen going from a 3 point basis to 2.

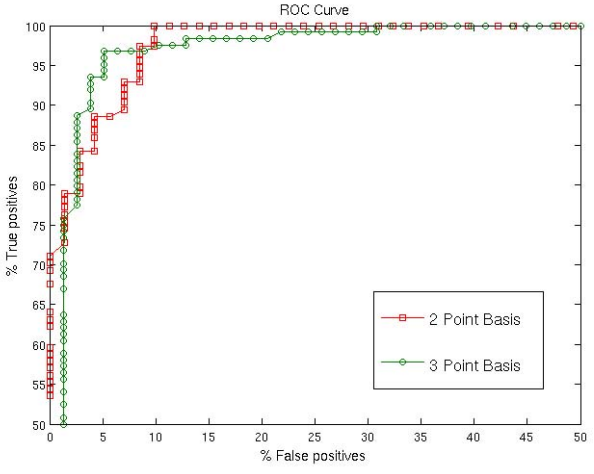

Fig. 7. Comparison between 2 and 3 point basis

Detection is reliant upon a good edge detection and therefore Chamfer map, however, edge information varies with lighting conditions, scale and subject. For reliable recognition, a suitable edge threshold is required for the given subject. An optimal threshold (OT) classifier was therefore constructed in a similar way to the previous classifiers, this is then used to alter the edge threshold at run time. This allows continuous parameter selection and thus our system is more robust to subject change and lighting conditions. The OT classifier is build using positives examples from different expressions with an optimal threshold selected manually. The negative examples are the same expressions with extremely low (high noise ratio) edge thresholds. This allows boosting to select features which are consistent across all expressions at an optimal edge threshold and more importantly negative features consistent across expressions at low edge thresholds. Since the features of the face provide strong edge information across a range of edge thresholds the OT classifier was predominantly constructed from negative features which are consistent only at low edge thresholds. At runtime the response of the OT classifier will peak at a suitable threshold which can then be used with the other classifiers.

6 Conclusions and Future Work

In this paper a novel automatic facial expression classifier is presented. Unlike other popular methods using Gabor wavelets, we have developed a real time system. For a 6 class (Joy, Surprise, Sadness, Fear, Anger and Disgust) system a recognition rate of 85% is achieved. Recognition is done on a frame by frame basis. As suggested in the literture [7], ensemble methods can often outperform single classifiers. In our experiments, the ensemble classifier approach provided an increase of up to 27% in recognition rates.

Bassili [14] demonstrated how temporal information can improve recognition rates in humans. Some faces are often falsely read as expressing a particular emotion, even if their expression is neutral, because their proportions are naturally similar to those that another face would temporarily assume when emoting. Temporal information can overcome this problem by modeling the motion of the facial features. Future work will incorporate temporal information into the current approach.

Acknowledgement

This work is partly funded by EPSRC through grant LILiR (EP/E027946) to the University of Surrey.

References

1. Alessandro, L.F.: A neural network facial expression recognition system using unsupervised local processing. In: ISPA 2001. 2nd international symposium on image and signal processing and analysis, Pula, CROATIE, pp. 628–632 (2001)
2. Barrow, H.G., Tenenbaum, J.M., Bolles, R.C., Wolf, H.C.: Parametric correspondence and chamfer matching: Two new techniques for image matching. In: DARPA 1977, pp. 21–27 (1977)
3. Bartlett, M., Littlewort, G., Fasel, I., Movellan, J.: Real time face detection and facial expression recognition: Development and application to human-computer interaction (2003)
4. Gong, S., McOwan, P., Shan, C.: Conditional mutual information based boosting for facial expression recognition. In: British Machine Vision Conference (2005)
5. Choudhury, T., Pentland, A.: Motion field histograms for robust modeling of facial expressions. In: ICPR 2000. Proceedings of the International Conference on Pattern Recognition (2000)
6. Dailey, M.N., Cottrell, G.W.: Pca = gabor for expression recognition. Technical report, La Jolla, CA, USA (1999)
7. Dietterich, T.G.: Ensemble methods in machine learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000)
8. Ekman, P., Friesen, W.V., Hager, J.C.: Facial action coding system. Palo Alto, CA, USA (1978)
9. Essa, I.A., Pentland, A.: Facial expression recognition using a dynamic model and motion energy. In: ICCV, pp. 360–367 (1995)

10. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: International Conference on Machine Learning, pp. 148–156 (1996)
11. OpenCV User Group. OpenCV Library Wiki (2006), http://opencvlibrary.sourceforge.net
12. Kanade, T., Tian, Y., Cohn, J.F.: Comprehensive database for facial expression analysis. In: FG 2000. Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000, p. 46. IEEE Computer Society Press, Washington, DC, USA (2000)
13. Mahamud, S., Hebert, M., Shi, J.: Object recognition using boosted discriminants. In: CVPR 2001. IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, IEEE Computer Society Press, Los Alamitos (December 2001)
14. Bassili, J.N.: Facial motion in the perception of faces and of emotional expression.
15. Saatci, Y., Town, C.: Cascaded classification of gender and facial expression using active appearance models. In: FGR 2006. Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, pp. 393–400. IEEE Computer Society Press, Washington, DC, USA (2006)
16. Shotton, J., Blake, A., Cipolla, R.: Contour-based learning for object detection. In: ICCV 2005. Proceedings of the Tenth IEEE International Conference on Computer Vision, vol. 1, pp. 503–510. IEEE Computer Society Press, Washington, DC, USA (2005)
17. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vision 57(2), 137–154 (2004)
18. Wang, Y., Ai, H., Wu, B., Huang, C.: Real time facial expression recognition with adaboost. In: ICPR, vol. 03, pp. 926–929 (2004)
19. Whitehill, J., Omlin, C.W.: Haar features for facs au recognition. In: FGR 2006. Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, pp. 97–101. IEEE Computer Society Press, Washington, DC, USA (2006)
20. Yacoob, Y., Davis, L.S.: Recognizing human facial expressions from long image sequences using optical flow. IEEE Trans. Pattern Anal. Mach. Intell. 18(6), 636–642 (1996)
21. Zhang, Z., Lyons, M., Schuster, M., Akamatsu, S.: Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In: FG 1998. Proceedings of the 3rd. International Conference on Face & Gesture Recognition, p. 454. IEEE Computer Society Press, Washington, DC, USA (1998)