Facial Expression Recognition Using Spatiotemporal Boosted Discriminatory Classifiers

Stephen Moore, Eng Jon Ong, and Richard Bowden

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

Abstract. This paper introduces a novel approach to facial expression recognition in video sequences. Low cost contour features are introduced to effectively describe the salient features of the face. Temporalboost is used to build classifiers which allow temporal information to be utilized for more robust recognition. Weak classifiers are formed by assembling edge fragments with chamfer scores. Detection is efficient as weak classifiers are evaluated using an efficient look up to a chamfer image. An ensemble framework is presented with all-pairs binary classifiers. An error correcting support vector machine (SVM) is utilized for final classification. The results of this research is a 6 class classifier (*joy*, *surprise*, *fear*, *sadness*, *anger* and *disgust*) with recognition results of up to 95%. Extensive experiments on the Cohn-kanade database illustrate that this approach is effective for facial exression analysis.

1 Introduction

The objective of this work is to exploit temporal information to build boosted classifiers for frontal facial expression recognition in video sequences. Facial expression recognition is a difficult task due to the natural variation in appearance of subjects. Such variation include ethnicity, age, facial hair, occlusion, pose and lighting. Many fields benefit from accurate facial expression recognition including behavioral science, security, communication and education. This paper presents an approach that relies on temporal boosted discriminatory classifiers based upon contour information. Contours are largely invariant to lighting and as will be shown, provide efficient classifiers using chamfer matching.

Cross cultural studies in Psychology signify a correlation between base emotions and facial expressions [8]. Current facial expression recognition systems highlight this observation by classifying a set of prototypical emotions such as *joy*, *surprise*, *fear*, *sadness*, *anger* and *disgust* [16,13]. Two common approaches for feature extraction for facial expression recognition are geometric based and appearance based methods [20]. Geometric features exploit shape and location information of facial components. Appearance based features capture the appearance change of the face (including wrinkles, bulges and furrows) and are

A. Campilho and M. Kamel (Eds.): ICIAR 2010, Part I, LNCS 6111, pp. 405–414, 2010.

extracted by image filters applied to the face or sub regions of the face. Geometric features are sensitive to noise and usually require reliable and accurate facial feature detection and tracking. However, appearance based features are less reliant on initialization, do not suffer from tracking errors and can encode changes in skin texture that are important for facial expression recognition. This paper investigates appearance based features based upon contour information. Humans have the ability to recognize facial expressions from a simplified line drawing or cartoon of the face. Sufficient information must therefore be present in this simplified representation for a computer to recognize facial expressions. Using only contour information provides important advantages as it offers some invariance to lighting and reduces the complexity of the problem.

Temporal information is incorporated by using a boosting framework [18] with the potential to develop weak classifiers by utilizing previous frames response in evaluating the current frame. This algorithm also incorporates temporal consistency of the data to facilitate recognition. We investigate the use of an ensemble classifier design to improve recognition. For final classification an error correcting SVM is used.

This rest of this paper is organized as follows. Related work is presented in section 2. Section 3 explains the methodology of this research. Section 4 outlines the data and experiments used to evaluate this research. Finally conclusions are presented in section 5.

2 Related Work

Facial expression recognition can be performed by using features from one image or by considering information from a image sequence. Research in psychology shows image sequences provide more accurate information than single frames. Bassili [2] conducted experiments showing that human facial expression recognition is superior when dynamic images are available. 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 face. Utilizing temporal information can translate to more robust and accurate classification when compared with static classifiers.

Hidden Markov Models (HMM) are frequently applied to spatio temporal facial expression recognition as they model the dynamics of expressions. Oliver et al. introduce two dimensional blob features to track mouth motion and uses HMMs to classify facial expressions [14]. Cohen et al. [5] proposed a multilevel HMM that uses the state sequence of independent HMMs to segment and recognize facial expressions. However flow estimates are easily disturbed by changes in illumination and non rigid motion.

Another way to capture temporal information is to map images to low dimensional manifolds for different expressions. Chang et al. [4] created a manifold from sparse 2d points. Shan et al. [16] used local binary pattern features for the

whole face to create manifolds of facial expressions and used a bayesian temporal model for facial expression recognition.

Zhao and Pietikainen introduced a novel approach for recognizing dynamic texture for classifying facial expressions [21]. Dynamic texture is an extension of texture into the temporal domain. Volume local binary patterns were proposed to capture appearance and motion. Petridis and Pantic investigated audio and visual temporal features for laughter detection [15]. Features were extracted for each frame in a temporal window. Mean, standard deviation and polynomials were calculated over the temporal window. Sheerman-Chase et al. [17] used similar temporal features for detection of non verbal facial displays. Yang et al. [19] introduces dynamic binary patterns based on harr like features to represent the dynamics of facial expressions.

Moore and Bowden [13] introduced edge and chamfer features for static facial expression recognition. Adaboost was used to learn discriminatory features and competitive results were obtained. In this research we apply the same features in a temporal framework for facial expression recognition in video sequences. Boosting techniques rarely utilize temporal information for classification. Recently Smith et al. introduced temporalboost which introduced temporal consistency in a boosting framework [18]. The algorithm averages weak classifiers from previous frames while the classification error decreases. In this research we investigate how temporal information in facial expressions can be utilized using temporalboost.

3 Methodology

3.1 Overview

The following section introduces how our facial expression classifier works, illustrated in figure 1. Images are annotated (two eyes and the tip of the nose) to allow features to be transformed to a reference co-ordinate system. The canny edge algorithm is used to create edge maps of all images. From each edge map, coherent edge fragments are extracted from the area in and around the face. A classifier bank is built containing all the edge fragments. Weak classifiers are created by combining edge fragments with a chamfer score. Temporalboost learns the optimal subset of features from the classifier bank and forms a strong classifier. Previous studies have shown a performance increase when ensemble classifiers are used [13], we adopt a similar approach resulting in 15 binary classifiers. These binary classifiers are used as input to a error correcting SVM for final classification.

3.2 Feature Extraction

Images are manually annotated to identify the two eyes and the tip of the nose, to form a 3-point basis. A 3-point basis is sufficient to align examples as only frontal faces are considered. Most approaches to frontal facial expression recognition only consider a 2 point basis (the two eyes), however head movements

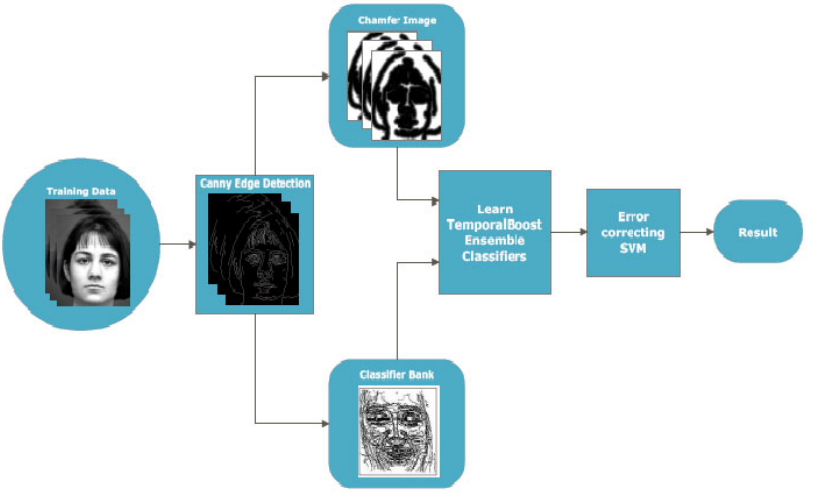

Fig. 1. Overview of Facial Expression Recognition System

are influenced by our emotions [12] so a 3 point basis as a reference co-ordinate frame is more tolerant to variations in head pose. 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.

The canny edge detection algorithm is used to extract edges [3]. In face images, edges characterize the boundaries of salient facial features as well as facial deformation due to facial expressions. First the detector smooths the image to eliminate noise. Next the sobel operator performs spatial gradient measurement on an image. The image is then scanned along the image gradient direction and if pixels are not part of the local maxima they are set to zero. This subdues image information that is not part of local maxima and is called non maximum suppression. A threshold is then used to evaluate if the magnitudes are sufficient to be classified as an edge e. $E = \{e\}$, where E is the edge map of an image. This threshold is selected manually for the dataset so salient features of the face are visually coherent in the edge maps. Following edge detection, connected component analysis is performed. From each edge component, short edge fragments T are extracted with variable lengths (based on heuristics of the face).

3.3 Chamfer Image

To measure support for any single edge feature over a training set we need some method for measuring the edge strength along that feature in a image. This can be computed efficiently using Chamfer matching. Chamfer matching was first introduced by Barrow et al. [1]. It is a registration technique whereby a drawing consisting of a binary set of features (contour segments) is matched to an image. A distance transform converts a binary image, which consists of

feature and non-feature pixels, into an image where each pixel value denotes the distance to the nearest feature pixel. Thus similarity between two shapes can be measured using their chamfer distance. The matching of a template with the chamfer image rather than the original edge image gives an advantage, as the resulting measure will be smoother as a function of the template transformation parameters [10].

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

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

A chamfer score is evaluated for each contour fragment T, where $T = \{t\}$:

$$d_{cham}^{(T)}(DT_E) = \frac{1}{N} \sum_{t \in T} DT_E(t) \tag{2}$$

where N is the number of edge points in T. This gives the Chamfer score as a mean distance between feature T and the chamfer image DT_E. The function $d_{cham}^{(T)}(DT_E)$ is an efficient lookup to the chamfer image for all classifiers. An example of a chamfer image is shown in figure 1.

3.4 Temporalboost

Boosting is a machine learning algorithm that produces a very accurate strong classifier, by combining weak classifiers in linear combination. Adaboost was introduced by Freund and Schapire [9] and has been successfully applied to static facial expression recognition [13]. Smith et al. [18] introduced Temporalboost, a boosting algorithm that introduces temporal consistency, by averaging weak classifiers sequentially. This allows weak classifiers to utilize information from previous frame when evaluating the current frame.

Temporalboost is an extension of adaboost. Like adaboost, 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. However, temporalboost modifies adaboost by allowing the best weak classifier to use previous frames responses if the overall classification error is decreased. This is achieved by using an OR operation and an AND operation for the previous t responses. The OR operation will respond positively if any of the previous t frames were classified as positive. This can allow for more true positives at the cost of false positives [18]. The AND operation will respond positively if all the previous t frames were classified as positive. This operation will decrease the number of false positives at the cost of true positives [18]. For example, if a feature classifies an event correctly for the previous t frames, but missclassifies the current frame, then

temporalboost allows the current frame to be classified correctly by using the previous t responses if the overall classification error is decreased. Both operations are performed for each iteration and the operation with the minimum classification error is selected. These operations allow temporal smoothing to be part of the boosting framework. The temporal window t starts at *0* and is expanded as long as the overall classification error for the current weights is decreased. The temporalboost algorithm tries to separate training examples by selecting the best weak classifier $h_j(x)$ that distinguishes between the positive and negative training examples. A weak classifier thus consists of a feature (f_j), a threshold θ_j and a parity (p_j) indicating the direction of the inequality sign.

$$h_j(x) = \left\{ \begin{array}{l} 1 \; if \; p_j f_j < p_j \theta_j \\ 0 \; otherwise \end{array} \right\} \tag{3}$$

Where

$$f_j = d^{(T)}_{cham}(DT_E) \tag{4}$$

θ_j is the weak classifier threshold. Setting a fixed threshold requires a priori knowledge of the feature space, an optimal θ_j is found through an exhaustive search for each weak classifier. This allows the learning algorithm to select a set of weak classifiers with low thresholds that are extremely precise allowing little deviation. Also, weak classifiers with high thresholds, which allows consistent deformation of the facial features can be selected. This increases the performance, but as will be seen, does not result in over fitting of the data. An image can have up to 500 features. Thus, over the training set, many hundreds of thousands of features are evaluated during the learning algorithm.

3.5 Ensemble Architecture

Dietterich argued that ensemble methods can often perform better than a single classifier [7]. Temporalboost is a binary classifier. There are several ways to partition the classification task into binary decisions. The simplest way is to train 1 against all. Another approach is to train all possible combinations of classes (1:1). [13] showed an all pairs ensemble (1:1) outperformed the 1 against all method. We adopt an all pairs ensemble framework, which for n classes can be broken into $\frac{n(n-1)}{2}$ binary classifiers respectively. This allows each expression to be exclusively boosted against every other expression.

An error correcting SVM is used for final classification. An SVM classifier is adopted here since it is a well understood classification technique that has been demonstrated to be effective in facial expression recognition. An SVM takes a feature vector as input in an n-dimensional space and constructs a separating hyperplane in that space, one which will maximize the margin between the positive and negative sets. The better the hyperplane, the larger the distance to the neighboring points from both classes. SVMs are usually binary classifiers, here we used a multi class SVM [6] which uses a one against all approach to solve the 6-class problem. The output from the temporalboost classifiers forms the input vector for the SVM. The SVM is trained using noisy training data by randomly perturbing the 3 point basis for each sequence.

4 Expression Classification

The Cohn-kanade facial expression database [11] 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 were captured as 640 x 480 png images. In each sequence, the subject started with a neutral expression and the sequence ends with the peak of the expression. In total 365 video sequences were chosen from the database (over 4,000 images). The only criteria was that the video sequence represented one of the prototypical expressions. This database is encoded using the Facial Action Coding System (FACS). A movement of one or more muscles of the face is called an action unit (AU) and all expressions can be described by a combination of one or more of 46 AU's. Each image has a FACS code and from this code, images are grouped into different expression categories.

Experiments were carried out using 5-fold cross validation with training and test sets divided 80-20. Due to the large number of features and training images we limited the number of boosting iterations to 500. In general about 20-30% of the weak classifiers selected have temporal paramters. Of the temporal weak classifiers selected, the majority use the OR operation. This reflects the fact that the data is not very temporally consisent and thus features using the AND operation don't minimize the classification error. Due to space restrictions the following discussion will focus on the joy ensemble classifiers (similar observation as discussed below are present for other expressions). Figure 2 shows the receiver operating characteristic (ROC) curves for all the *joy* ensemble classifiers. The more accurate classifiers are *joy* against *surprise* and *joy* against *anger*. This is as expected as the facial deformation due to the *joy* expression is very distinct from expressions *surprise* and *anger*. The worst performance is achieved with the *joy* against *disgust* classifier. This is due to the close proximity of the distinctive features (deformation around the cheek area) for these expressions.

Figure 3 visualizes the features which contribute to the classification of expressions. In general we can see that the contour around the edge of the mouth and the contour around the cheek are used to classify the *joy* expression. However as can be seen, depending on the negative expression different areas of the mouth and cheek contribute more to classification. For example in figure 3 pictures A, C and E show features from the corners of the mouth and the cheeks are prominent. Expressions *surprise*, *sad* and *anger* deform the face very differently to *joy* and thus all the deformation of the *joy* expression is captured in these classifiers. While pictures B and D show the importance of the corners of the mouth and not the deformation around the cheek. This is because the expressions *fear* and *disgust* deform the area around the cheek in a similar fashion to the *joy* expression. Another interesting observation in image B is the amount of noise. This finding can be explained by the fact that the expressions *joy* and *fear* are often difficult to disambiguate.

Table 1 shows the confusion matrix for 5 fold cross validation on the Cohn-kanade database. An overall recognition rate of 86.1% is achieved. From the

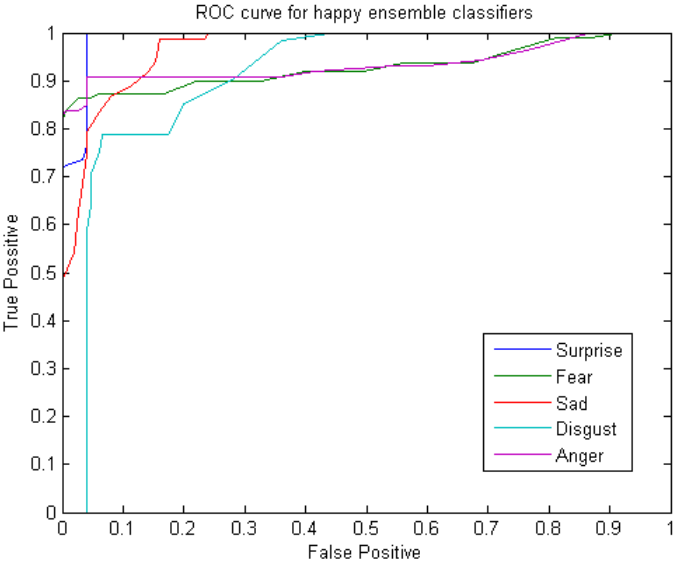

Fig. 2. Roc curves for joy ensemble classifiers

Fig. 3. Visualization of weak classifiers which contributed to classification. From left to right images represent: A) joy against surprise, B) joy against fear, C) joy against sad, D) joy against disgust and E) joy against anger.

results it is apparent that the more subtle expressions (*anger*, *fear* and *sad*) are outperformed by expressions with a large deformation (*joy*, *surprise*,*disgust*). Subtle changes in appearance are difficult to distinguish when using one reference co-ordinate frame due to the variability across subjects. Also it must be noted that the combination of contour and chamfer matching is varient to scale and rotation. Thus subtle expression are harder to disambiguate using these features.

The lowest recognition rate was for the *fear* classifier. Most confusion occurs between expressions *disgust* and *anger* due to similar deformation around the eyebrows. Also confusion occurs between *fear* and *sad* and between *sad* and *anger* classes respectively. In particular *sad* and *anger* expressions have little deformation when compared to expressions *surprise* or *joy*. This in itself could contribute to the confusion as a lack of distinct features makes the learning of strong classifiers more difficult. Also when posing a *sad* expression subjects can

Table 1. Confusion matrix for 5-fold cross validation on Cohn-kanade database

	Joy	**Surprise**	**Fear**	**Sad**	**Disgust**	**Anger**
Joy	**93.92**	0	2.94	1.67	1.47	0
Surprise	0	**95.09**	0	1.79	3.12	0
Fear	5.63	0	**75.55**	9.34	3.71	5.77
Sad	0	0	6.28	**85.36**	0	8.36
Disgust	0	0	2.78	0	**91.32**	5.9
Anger	0	0	0	9.32	15.1	**75.58**

exaggerate the expression and the mouth can have a similar appearance to the *fear* expression.

5 Conclusions

This paper presents a novel approach to frontal facial expression recognition in video sequences. Unlike other popular methods like Gabor wavelets, we present a fast efficient system that yields a recognition rate of 86.1%. Recognition is achieved on a frame by frame basis but classifiers use feature responces from previous frames to evaluate the current frame. An ensemble framework is presented which includes an all pairs architecture with an error correcting SVM for final classification. Competitive results were achieved on the Cohn-kanade database for 6 basic expressions. Expression with large deformation of the face achieved the best results with *surpise* achieving over 95% accuracy.

Acknowledgement

This work has been supported by the EPSRC project LILiR and by the FP7 project DICTASIGN (FP7/2007-2013) under grant agreement n 231135.

References

1. Barrow, H.G., Tenenbaum, J.M., Bolles, R.C., Wolf, H.C.: Parametric correspondence and chamfer matching: two new techniques for image matching. In: IJCAI 1977: Proceedings of the 5th International Joint Conference on Artificial Intelligence, pp. 659–663. Morgan Kaufmann Publishers Inc., San Francisco (1977)
2. Bassili, J.N.: Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. Journal of personality and social psychology 37(11), 2049–2058 (1979)
3. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. 8(6), 679–698 (1986)
4. Chang, Y., Hu, C., Turk, M.: Probabilistic expression analysis on manifolds. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 520–527 (2004)
5. Cohen, I., Garg, A., Huang, T.S.: Emotion recognition from facial expressions using multilevel hmm. In: Neural Information Processing Systems (2000)

6. Crammer, K., Singer, Y., Cristianini, N., Shawe-taylor, J., Williamson, B.: On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research 2, 2001 (2001)
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.: Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 124–129 (1971)
9. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: International Conference on Machine Learning, pp. 148–156 (1996)
10. Gavrila, D.: Pedestrian detection from a moving vehicle. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 37–49. Springer, Heidelberg (2000)
11. 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, Washington, DC, USA, p. 46. IEEE Computer Society, Los Alamitos (2000)
12. Mignault, A., Chaudhuri, A.: The many faces of a neutral face: Head tilt and perception of dominance and emotion. Journal of Nonverbal Behavior 27(2), 111–132 (2003)
13. Moore, S., Bowden, R.: Automatic facial expression recognition using boosted discriminatory classifiers. In: Zhou, S.K., Zhao, W., Tang, X., Gong, S. (eds.) AMFG 2007. LNCS, vol. 4778, pp. 71–83. Springer, Heidelberg (2007)
14. Oliver, N., Pentland, A., Brard, F.: Lafter: Lips and face real time tracker with facial expression recognition. In: Proc. CVPR, pp. 123–129 (1997)
15. Petridis, S., Pantic, M.: Audiovisual laughter detection based on temporal features. In: IMCI 2008: Proceedings of the 10th International Conference on Multimodal Interfaces, pp. 37–44. ACM, New York (2008)
16. Shan, C.F., Gong, S.G., McOwan, P.W.: Dynamic facial expression recognition using a bayesian temporal manifold model. In: BMVC 2006, pp. 297–306 (2006)
17. Sheerman-Chase, T., Ong, E.-J., Bowden, R.: Feature selection of facial displays for detection of non verbal communication in natural conversation. In: IEEE International Workshop on Human-Computer Interaction, Kyoto (October 2009)
18. Smith, P., da Vitoria Lobo, N., Shah, M.: Temporalboost for event recognition. In: ICCV 2005: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV 2005), Washington, DC, USA, vol. 1, pp. 733–740. IEEE Computer Society, Los Alamitos (2005)
19. Yang, P., Liu, Q.S., Cui, X.Y., Metaxas, D.N.: Facial expression recognition using encoded dynamic features, pp. 1–8 (2008)
20. Tian, Y., Kanade, T., Cohn, J.: Facial expression analysis. In: Handbook of Face Recognition, ch. 11, pp. 247–275. Springer, Heidelberg (2005)
21. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. 29(6), 915–928 (2007)