Facial Action Recognition using sparse appearance descriptors and their pyramid representations

Bihan Jiang[1], Michel F. Valstar[1], and Maja Pantic[1,2]

[1] Department of Computing, Imperial College London, UK
{bi.jiang09, michel.valstar, m.pantic}@imperial.ac.uk
[2] EEMCS, University of Twente, Netherlands
PanticM@cs.utwente.nl

Abstract. Most existing work on automatic analysis of facial expressions has focused on a small set of prototypic emotional facial expressions such as fear, happiness, and surprise. The system proposed here enables detection of a much larger range of facial behaviour by detecting facial muscle actions (action units, AUs). It automatically detect all 9 upper face AUs using local apperance descriptors. Meanwhile, the merits of the family of local binary pattern descriptors are investigated. We compare Local Binary Patterns, Local Phase Quantisation, Pyramid Local Binary Pattern, as well as our proposed descriptors Block-based Pyramid Local Binary Pattern and Block-based Pyramid Local Phase Quantisation for AU detection. Results show that our proposed descriptor Block-based pyramid Local Binary Pattern outperforms all other tested methods for the problem of FACS Action Unit analysis and the systems that utilise pyramid representation outperform those that use basic appearance descriptors.

1 Introduction

One limitation of the majority of existing facial expression recognition methods is that they focus on a small set of prototypic emotional facial expressions, specifically fear, sadness, happiness, anger, disgust, and surprise. Yet, these six basic emotion categories form only a subset of the total range of possible facial displays and the categorisation of facial expressions can therefore be forced and unnatural. The Facial Action Coding System (FACS) is the best known and most commonly used system developed for human observers to describe facial activities. The coding system defines atomic facial muscle actions called Action Units (AUs). With FACS, every possible facial expression (emotional or otherwise) can be described as a combination of AUs. For instance, the expression of happiness contains AU6 and AU12, while the expression of sadness contains AU1, AU4 and AU15.

Deriving an effective facial representation from images is an essential step for successful facial expression recognition. Traditionally the feature extraction

Andrew V. Jones (Ed.): Proceedings of the 1st Imperial College Student Workshop (ICCSW '11), pp. 39–45. September 29th–30th 2011, London UK.

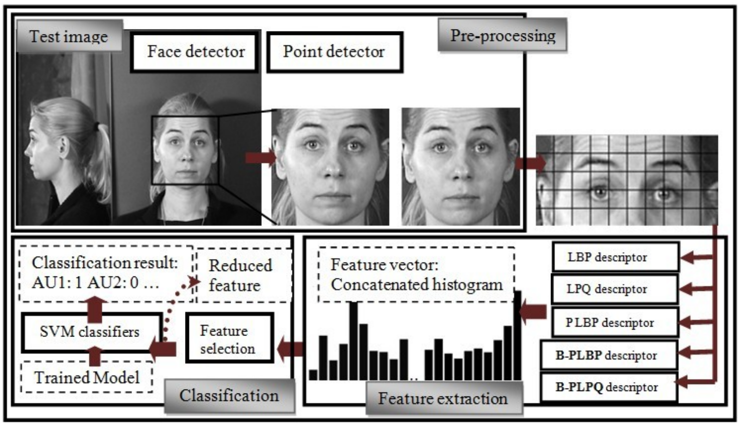

Fig. 1. The outline of our proposed system

approaches may be divided into two streams: geometric feature-based methods and appearance-based methods. Geometric feature based methods employ the geometrical properties of a face such as the positions of facial points relative to each other, the distances between pairs of points or the velocities of separate facial points. For a method using appearance features, the changes in image texture such as those created by wrinkles, bulges, and changes in feature shapes are captured.

Our key contributions are three-fold. First, we propose the novel appearance feature descriptors Block-based Pyramid Local Binary Pattern (B-PLBP) and Block-based Pyramid Local Phase Quantisation (P-BLPQ). Secondly, the proposed appearance descriptor B-PLBP and B-PLPQ are applied to the problem of FACS AU analysis. Finally, the applicability of different SVM kernels for histogram-based features has been studied. The experimental results show that our novel descriptor B-PLBP outperforms the three other methods for FACS AU analysis in terms of recognition accuracy.

The remainder of this paper is organised as follows. Section 2 briefly describes the methodologies used in this work. It introduced the basic principle of static appearance descriptors LBP, LPQ, PLBP and our proposed extensions B-PLBP and B-PLPQ, the training datasets used in our experiments, the classification technique used in this work and the different kernels tested. The evaluation procedures and test results are given in Section 3. Section 4 provides the conclusions of our research.

2 Methodologies

Fig. 1 shows an overview of the proposed system. In order to detect the upper face AUs, we use 9 SVM classifiers, one for each AU, which are trained on a subset of the most informative spatiotemporal features selected by GentleBoost. To extract these appearance features, we first find the face in the input static image using an adapted version of the Viola and Jones face detector. Next the detected face images are registered to remove head rotations and scale variations by using the OpenCV implementation of an object detector to locate the eyes. Based on that, the face image is scaled to make the distance between the eye locations 100 pixels, and then cropped to be 200 by 200 pixels. After that, the registered image is divided into small blocks and the LBP, LPQ, PLBP, B-PLBP and B-PLPQ features are extracted. The histograms from all blocks are concatenated as a feature vector to represent the corresponding face image.

2.1 Local Appearance Descriptors

Method 1. Local Binary Patterns (LBP) were first introduced by Ojala et al. in [4], and proved to be a powerful means of texture description. By thresholding a 3×3 neighbourhood of each pixel with respect to the centre value, the operator labels each pixel. Considering the 8-bit result to be the binary representation of a decimal number, a 256-bin histogram of the LBP labels computed over a region is used as a texture descriptor. This has been successfully applied to face recognition by Ahonen et al.[1]. They proposed to divide face images into m local regions, from which LBP histograms can be extracted, and then concatenate them into a single,spatially enhanced feature histogram. The resulting histogram encodes both the local texture and global shape of face images. This version is what we adopted in our work. Readers are kindly asked to refer to [4, 1] for details.

Method 2. Local Phase Quantisation (LPQ) was originally proposed by Ojansivu and Heikkila as a texture descriptor that is robust to image blurring [5]. The descriptor uses local phase information extracted using the 2-D DFT or, more precisely, a short-term Fourier transform (STFT) computed over a rectangular M-by-M neighbourhood N_x at each pixel position $\mathbf{x}$ of the image $f(\mathbf{x})$ defined by

$$F(\mathbf{u},\mathbf{x}) = \sum_{\mathbf{y}\in N_x} f(\mathbf{x}\text{-}\mathbf{y})e^{-j2\pi\mathbf{u}^T\mathbf{y}} = \mathbf{w}_\mathbf{u}^T\mathbf{f}_\mathbf{x} \tag{1}$$

where $\mathbf{w}_\mathbf{u}$ is the basis vector of the 2-D DFT at frequency $\mathbf{u}$, and $\mathbf{f}_\mathbf{x}$ is the vector containing all M^2 samples from N_x.

The phase information in the Fourier coefficients is recorded by examining the signs of the real and imaginary parts of each component in F_x. The resulting eight bit binary coefficients $g_j(x)$ are represented as integers using binary coding. As a result, a histogram of these values from all positions is composed and used as a 256-dimensional feature vector in classification. Similar to LBP, we use a block version of LPQ which has shown promising performance in [3]. For more details, please refer to [5, 3].

Method 3. Qian et.al [6] extended the conventional LBP to the pyramid transform domain named **Pyramid Local Binary Pattern (PLBP)**. By cascading the LBP information of hierarchical spatial pyramids, PLBP takes texture resolution variations into account. They comprehensively compared PLBP with other LBP extensions for texture classification and claimed that PLBP is with satisfactory performances and with low computational cost. However, a histogram computed over the whole image represents only the global distribution of the patterns thus the local information has been ignored. On the other hand, some researchers are critical of Ahonen's approach, suggesting that the subregions are not necessarily well aligned with facial features and the resulting facial description depends on the chosen window size and the position of these subregions [2]. These problems were reflected in our results (see Fig.5).

Motivated by these ideas, we propose two novel descriptors **B-PLBP** and **B-PLPQ** which capture pixel-level, region-level and structure-level information for face representation. The face image is represented in an image pyramid by different spatial resolutions. Each pixel in the higher spatial pyramid levels is obtained by down sampling from its adjacent low-pass filtered high resolution image. Hence in the low resolution images, a pixel corresponds to a region in its high-resolution equivalently. For each pyramid level, the image are divided into regions. The region sizes remain constant. The dense appearance descriptor features extracted from each region, and each level of the pyramid, are concatenated into a single, spatially enhanced feature histogram. As shown in Fig.2, the blocks in each level encodes different spatial information. In our experiments, a three level pyramid model and a region size of 25 × 25 pixels is used.

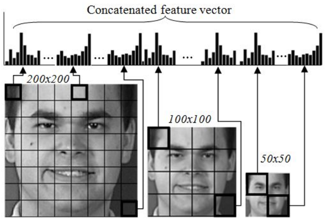

Fig. 2. The block-based pyramid representation

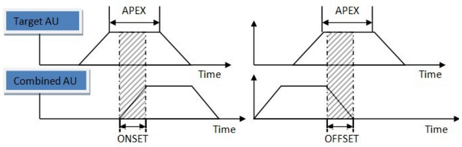

Fig. 3. The criterion of static data selection. The shaded areas are included in the dataset

2.2 Data Collection

In this work, the efficiency of the discussed descriptors are tested based on dataset collected from the MMI Facial Expression Database (MMI database [8]). The MMI database is a fully web-searchable collection of visual and audio-visual recordings of subjects displaying facial expressions which are FACS annotated. It includes 69 different subjects of both sexes (44 female), ranging in age from 19 to 62, having either a European, African, Asian, Caribbean or South American ethnic background. All fully FACS-coded recordings show facial expressions that are posed, and it is these data which will be used in this work.

In [3], the authors proposed a **heuristic approach** to select data for training. It is noted that when more than one AU is activated, facial actions can appear very different from when they occur in isolation. For example, AU1 and AU2 pull the brow up, whereas AU4 pulls the brows together and down using primarily the corrugators muscle at the bridge of the nose. The appearance of AU4 changes dramatically depending on whether it occurs alone or in combination with AU1 and AU2. In order to capture the appearance of each action unit as fully as possible and thus build a richer data space, the heuristic approach takes in every video the first apex frames of each target AU, and all the apex frames where any other upper face AUs are in onset or offset (see Fig. 3). The shaded parts are the frames selected. However, AU combinations are not treated differently by the classifiers. In other words, each AU is recognised independently of all the others.

2.3 Classification

A previous successful technique to facial expression classification is Support Vector Machine (SVM). In this work, we adopted SVM as classifiers for AU detection. Given a training set of labelled examples $\{(x_i, y_i), i = 1, ..., l\}$, where $x_i \in R^n$ and $y_i \in \{1, -1\}$, a new test example x is classified by the following function:

$$f(x) = sgn(\sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x_i}, \mathbf{x}) + b) \tag{2}$$

where sgn function returns the sign of y, i.e. either 1 or -1, α_i are Lagrange multipliers of a dual optimisation problem that describe the separating hyperplane, $K()$ is a kernel function, and b is the threshold parameter of the hyperplane. Performing an implicit mapping of data into a higher dimensional feature space, which is defined by the kernel function, the training process is achieved by finding a linear separating hyper-plane with the maximal margin (M) to separate data in this higher dimensional space. The most popular kernels are linear, polynomial and Radial Basis Function (RBF). Recently, Maji et.al [7] proposed a histogram intersection kernel SVMs (IKSVMs). They also introduced a more efficient way to compute it. It is shown that IKSVM gives comparable accuracy while being 50× faster and require 200× less memory than the standard SVM implementation in their experiments. In this work, we evaluate the efficiency of these four kernels in our application.

3 Evaluation

3.1 Comparison Setup

We evaluated the four appearance descriptors on 442 videos taken from the MMI database. In order to compare different approaches, the same evaluation process is performed. As this is a user independent system for FACS Action Unit detection, the evaluation is done in a subject independent manner. Generalisation to new subjects is tested using 10-fold cross validation.

The performance measure used in this work is the area under the ROC curve. By using the signed distance of each sample to the SVM hyper-plane and varying a decision threshold, we plot the hit rate (true positives) against the false alarm rate (false positives). The area under this curve is equivalent to percent correct in a 2-alternative forced task (2AFC), which can be computed more efficiently.

3.2 Results

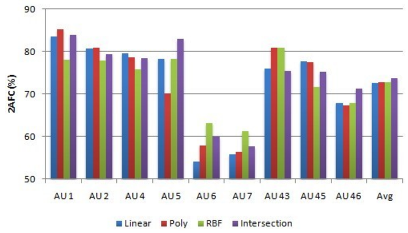

Fig. 4. Average 2AFC (%) based on different kernels for SVM

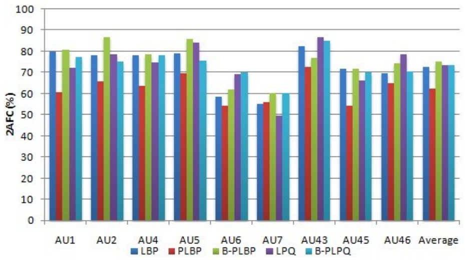

Fig. 5. The 2AFC (%) using LBP, PLBP, B-PLBP, LPQ and B-PLPQ based on MMI

Experiment 1. *Kernel functions:* Fig.4 shows the average 2AFC scores performed with B-PLBP based on different SVM kernel as we discussed in section 2.3. The LBP features and the proved best training data selection method, the heuristic approach, has been employed. For all the kernels, the parameters are optimised before training (refer to 3-A). In general, the best results are reached with the histogram intersection kernel. This is expected as all the features used in this work are histogram-based. For AU6 and AU7, which our detector poorly performed, RBF kernel gives the best result. This probably result from the fact that features that which capture subtle appearance changes, are non-linear separable.

Experiment 2. *Appearance descriptors:* Figure 5 presents the 10-fold cross-validation results using LBP, LPQ, PLBP, B-PLBP and B-PLPQ for 9 upper face AUs. Note that LBP and LPQ used here are block-based. To report the best performance of all systems, the heuristic approach and the histogram intersection kernel SVM are adopted in these experiments. In general, the block-based pyramid extension outperform their original version (LBP and LPQ). The importance is more clear for P-PLBP. We can see that, overall speaking, B-PLBP produces best results among these four descriptors and the PLBP performs worst.

The average 2AFC scores from B-PLBP is 12.8% higher than that for PLBP. The importance of local shape information for AU detection is again shown by our results. Compared to B-PLBP, the improvement of B-PLPQ is less obvious. This can probably be explained by the blur-invariant characteristic of the LPQ descriptor, which effectively negates the effect of the image pyramid.

4 Conclusions

We successfully implemented a robust and real-time AU detection system. We compared the appearance descriptors LBP, LPQ and their block-based pyramid extension B-PLBP and B-PLPQ. Results show that the systems based on LPQ generally achieve higher accuracy rate than LBP system, and that the systems that utilise pyramid representation outperform those that use basic appearance descriptors. Although the family of block-based pyramid descriptors are more computationally expensive than the basic ones, they attain a higher recognition performance. All in all, the experimental results clearly show that our proposed descriptor B-PLBP outperforms all other tested methods for the problem of FACS Action Unit analysis. Note that although we only applied the method to upper face AUs, the method can be readily used for all other AUs.

References

1. T. Ahonen, A. Hadid, and M. Pietikainen. Face recognition with local binary patterns. In *European Conference on Computer Vision*, pages 469–481, 2004.
2. D. Huang, C. Shan, and M. Ardabilian. Local binary pattern and its application to facial image analysis: A survey. *IEEE Trans. Systems, Man and Cybernetics, Part C*, 41:1–17, 2011.
3. B. Jiang, M. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11)*, Santa Barbara, CA, USA, March 2011.
4. T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on featured distribution. *Pattern Recognition*, 29(1):51–59, 1996.
5. V. Ojansivu and J. Heikkila. Blur insensitive texture classification using local phase quantization. In *In Proc. Int. Conf. on Image and Signal Processing*, volume 5099, pages 236–243, 2008.
6. X. Qian, X. Hua, P. Chen, and L.Ke. Plbp: An effective local binary patterns texture descriptor with pyramid representation. *Semi-Supervised Learning for Visual Content Analysis and Understanding*, 44(10):2502–2515, 2011.
7. J. M. S. Maji, A.C. Berg. Classication using intersection kernel support vector machines is efcient. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
8. M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. Int'l Conf. Language Resources and Evaluation, W'shop on EMOTION*, pages 65–70, 2010.