DISCOVERING THE BEST FEATURE EXTRACTION AND SELECTION ALGORITHMS FOR SPONTANEOUS FACIAL EXPRESSION RECOGNITION

Ligang Zhang, Dian Tjondronegoro and Vinod Chandran
Queensland University of Technology, Brisbane, 4000, Australia
Ligzhang@gmail.com, {dian, v.chandran}@qut.edu.au

ABSTRACT

Feature extraction and selection are critical processes in developing facial expression recognition (FER) systems. While many algorithms have been proposed for these processes, direct comparison between texture, geometry and their fusion, as well as between multiple selection algorithms has not been found for spontaneous FER. This paper addresses this issue by proposing a unified framework for a comparative study on the widely used texture (LBP, Gabor and SIFT) and geometric (FAP) features, using Adaboost, mRMR and SVM feature selection algorithms. Our experiments on the Feedtum and NVIE databases demonstrate the benefits of fusing geometric and texture features, where SIFT+FAP shows the best performance, while mRMR outperforms Adaboost and SVM. In terms of computational time, LBP and Gabor perform better than SIFT. The optimal combination of SIFT+FAP+mRMR also exhibits a state-of-the-art performance.

Index Terms— Facial expression recognition, performance comparison, feature selection, Gabor, SIFT

1. INTRODUCTION

Facial expression recognition (FER) is an important field for affective computing and supports many applications, including human computer interaction, video surveillance, and patient condition monitoring. Spontaneous FER is particularly important because it reflects the real reaction of humans mimicking real-world situations.

An FER system is normally composed of four main steps: face detection/tracking, feature extraction, feature selection, and emotion classification. Choosing suitable feature extraction and feature selection algorithms play the central roles in providing discriminative and robust information, particularly for spontaneous emotions which differ posed emotions in subtle ways in appearance and timing etc. [1]. However, it is difficult to isolate the subtle differences to check if a particular feature extraction algorithm will be more suited for spontaneous FER or not. In addition, spontaneous FER images are more likely to co-occur with changes in pose and illumination, and face movements etc. Accordingly, the features also should be extracted in a way that is robust to these changes.

There are many up-to-date studies that compare performance between different types of features [2], [3], [4], [5], [6], and selection algorithms [5], [7], [8], [9]. However, except for [4], [6], these studies have only benchmarked performance on posed emotions, rather than spontaneous ones. These existing studies have only adopted either texture or geometry features, whereas when combined, they can be complementary. Few of them [2], [5] compare different features and different feature selection algorithms in the same framework. To the best of our knowledge, direct comparison between texture, geometry and their fusion, as well as between multiple selection algorithms has not been found for spontaneous FER.

This paper presents a unified framework for performance comparisons between different types of features and feature selection algorithms, based on the standard Feedtum and NVIE databases of spontaneous emotions. Three texture features (SIFT, LBP, and Gabor) and one geometric feature (FAP) are used due to their wide adoption, state-of-the-art performance, and robustness to variations. These features are extracted around active shape model (ASM) points to enhance robustness to pose changes and face movements in spontaneous images. Three popular feature selection algorithms - Adaboost, minimal redundancy maximal relevance criterion (mRMR), norm-based support vector machine (SVM), are employed. An SVM is used for classifying facial expressions.

The rest of the paper is organized as follows. Section 2 summarizes related work. Section 3 describes the comparison framework, and Section 4 discusses experimental results. Finally, Section 5 outlines conclusions.

2. RELATED WORK

Various types of feature extraction and selection algorithms have been adopted in previous comparative studies on FER. For feature extraction, local binary patterns (LBP), Gabor and scale-invariant feature transform (SIFT) are the most frequently used algorithms. LBP features are the best performer amongst various features in many studies using both standard and real-world images [3], [4], [6], [10], [11]. Gabor features also represent the state-of-the-art FER performance [12]. SIFT has been reported as having better performance than LBP and HOG for multi-view FER [2]. All the three features have advantages of discriminating
feature extraction, robustness to illumination variations and noise, and insensitivity to a reasonable amount of changes in image scale and rotation. In addition, LBP is also known for computational simplicity and SIFT can provide robust matching across affine distortions. All these characteristics make the three features suitable for spontaneous FER. They have been successfully used in FER on real-world data [10].

For feature selection, Adaboost and mRMR are the two most widely used algorithms in previous FER work. Adaboost has shown significant improvements over other algorithms [10], [13], [14]. mRMR also has been demonstrated with a better performance than PCA, mutual information, and genetic algorithm [5], [9]. Norm-based SVM has also shown good performance in recent work on object and action recognition [15]. Thus, they will be adopted for the performance comparison.

3. COMPARISON FRAMEWORK

Fig. 1. Processing steps in the comparison framework.

Fig. 1 shows processing steps in the comparison framework. For an input image, the face is located using the Viola-Jones detector and 68 fiducial facial points are detected using a well-trained ASM. On one side, three most widely used texture features (LBP, SIFT and Gabor) are extracted around each of 53 interior points, and the vectors from all points are concatenated into a final vector for each type of feature. A subset of the most discriminative texture features is selected using three different algorithms, including Adaboost, mRMR and SVM. On the other side, the geometric feature vector is composed of 45 distances defined based on an ASM and FAPs. SVM with a radial basis function kernel is used for classifying facial expressions using the geometric and texture feature vectors individually and fused. Performance is evaluated on two public Feddum and NVIE databases with spontaneous emotions. The classified expressions include six basic emotions: ANger, DIsgust, FEar, HAppiness, SAdness, and SUprise, plus NEutral.

3.1. Face and fiducial point detection

For an input image, the face is detected using the widely used Viola-Jones detector. No pre-processing is conducted to simulate the real imaging situations. From the facial region, 68 facial fiducial points are detected using an ASM. ASM is known for its robustness in fitting and tracking fiducial points in human faces. To train the ASM, we collect 100 images from the Internet with different emotions and poses, ranging from -20 to 20 degree. Then 68 fiducial points as shown in Fig.2a are manually annotated with x and y locations. The trained ASM is anticipated to work well for detecting fiducial points in faces with normal face movements. It has been observed that the points in the face boundary (No.1 to 15 in Fig.2a) are not always accurately detected by the ASM due to face shape changes in different subjects and face movements (a case shown in Fig.2b). Further, the regions around these points contain background information and do not provide reliable features. Therefore, only 53 interior points are used to extract features.

3.2. Texture and geometric feature extraction

To maintain a reasonable degree of tolerance to face movements and pose changes, texture features are extracted around each of the 53 interior points. Features of all points are then combined into a vector. This method of extracting features helps to optimize FER performance [16]. Three descriptors, including LBP, Gabor and SIFT, are used due to their high performance in facial expression analysis.

LBP [17] labels each pixel in an image as a binary number by applying thresholds to neighborhood pixels with the center value, then accumulates the occurrence of different binary patterns, yielding a histogram as the texture descriptor of the image. Based on the setting [3], we collect uniform patterns LBP(4,2) with 59 labels from a 14×18 patch centered at each point, resulting in a histogram with 2,597 bins for all points.

Gabor features can be extracted by performing multi-scale and -orientation filters on an image. Following the common setting, we use five scales \( \lambda_m = 4 \times 2^{m-1} \); \( m = (1, ... 5) \) and eight orientations \( \theta_n = \pi(n - 1)/8 \) Gabor filters, which result in 40 Gabor magnitude coefficients for each point and a final feature vector with 2,120 elements.
SIFT [18] yields a kind of distinctive invariant features that are suitable for describing local features. Following the settings in [16], the SIFT descriptor is computed from the gradient vector histograms of the pixels in a 4x4 patch around each point. Given eight possible gradient orientations, each descriptor contains 128 elements, and the final feature vector contains 6,784 elements.

Geometric features include 43 distances between the 53 interior fiducial points. These distances are calculated based on facial animation parameters (FAPs) defined in the ISO MPEG-4 standard. Compared with facial movement vectors used in previous work, distance features have the merits of being robust to pose changes, and do not require the compensation of face movements. Therefore, they are suitable for the proposed framework working in spontaneous images. To allow a constant expression for arbitrary faces, FAP units (FAPUs) are defined as the fractions of distances between key points to scale FAPs. Details of these distances can be found in [Reference withheld for blind review].

3.3. Texture feature selection

Feature selection aims to choose a subset of the most discriminative features from the texture feature vector. Three algorithms, including Adaboost, mRMR and norm-based SVM, are selected for the comparison.

mRMR [19] selects a subset of features that jointly have the largest dependency on the ground truth class and the least redundancy among the features. The dependency and redundancy can be combined using the mutual information difference (MID) or quotient (MIQ). The MID is adopted here for its simplicity.

Adaboost aims to find a final hypothesis with a low error relative to a given weight distribution over the training examples. On each round, the weights are updated so that the weights of misclassified examples are increased, while those of correctly classified examples are decreased. This simple, yet effective selection criterion provides a quick solution to reduce the overall error. The multiple-emotion-class problem is handled by the one-against-all strategy.

Norm-based SVM treats the distances to the classification hyperplane as the weights of features, and updates these weights every round to reflect the importance of each feature. At each round, a one-against-all SVM is trained for each emotion class and the average weights over all classes are updated based on the trained SVM. The features with low weights are dropped out (i.e. setting the weights to 0).

4. EXPERIMENTS

4.1. Databases

The Feedtum database [20] was collected to assist researchers to investigate the effects of different facial expressions. It contains 399 video sequences from 18 subjects. Each subject performed all six basic plus neutral emotions three times, and each sequence starts and finishes with a neutral state. The database attempts to capture real emotions by probing the observed people’s natural reaction to video clips or still images, which may result in head movements, instead of asking them to pose for different emotions in one direction.

The natural visible and infrared facial expression (NVIE) database [21] is a newly developed comprehensive benchmark for facial expression analysis. The spontaneous expressions are induced by film clips deliberately selected from the Internet. There are 105, 111, 112 subjects in the spontaneous database under front, left and right illumination, respectively. During recording, all subjects are allowed to seat themselves comfortably and move the chair forward or backwards, resulting images having different sizes of faces and face movements. Spontaneous images are labeled by five students with six basic emotions. Fig. 3 shows image samples for seven emotions on the Feedtum and NVIE databases.

![Image samples for Feedtum (top) and NVIE databases.](image)

For the experiment, five images are evenly selected from each of all Feedtum sequences, starting from the peak frame. For NVIE, only spontaneous visible images with final evaluated annotations are used. It should be noted that a subset of spontaneous images were not provided with final annotations by the database’s authors. All the selected images go through the face and fiducial point detection, yielding 1,787 Feedtum images and 1,488 NVIE images. For NVIE, only 1,472 images are retained by excluding those with a near neutral emotion. Table 1 displays the distribution of the selected images over seven emotions.

| Table 1. Distribution of selected images over seven emotions |
|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Feedtum   | AN  | DI  | FE  | HA  | SA  | SU  | NE  |
| NVIE      | 265 | 270 | 232 | 255 | 270 | 235 | 260 |
|           | 229 | 266 | 211 | 315 | 236 | 215 |     |

4.2. Classification performance

Figs. 4 and 5 show the recognition accuracy of classifying facial expressions on the Feedtum and NVIE databases respectively using (a) three texture features, (b) three feature selection algorithms, and (c) three fusion strategies. The three fusion strategies include using texture alone, FAP-based distances alone, and their fusion. Note that the results of (a), (b), and (c) are based on Adaboost, SIFT, SIFT+Adaboost respectively due to their higher performance than the other features and algorithms.
Fig. 4a and 5a compare the performance of three textures when fused with FAP-based distances using the Adaboost algorithm. For both the databases, SIFT+FAP performs the best, which is followed by Gabor+FAP, whereas LBP+FAP shows the lowest performance. The overall performance obtained using SIFT+FAP is 1.8-6.1% and 2.4-7.8% higher than those obtained using Gabor+FAP and LBP+FAP, respectively on the Feedtum database. The corresponding performance improvements are 0.8-6.5% and 4.3-9.2% respectively on the NVIE database. When mRMR is used for feature selection, SIFT+FAP still is the best overall performer, while Gabor+FAP and LBP+FAP perform similarly. When SVM is employed for feature selection, LBP+FAP achieves a slightly better performance than SIFT+FAP and Gabor+FAP. Thus, for the two databases, SIFT+FAP has the best overall performance using three selection algorithms.

Figs. 4b and 5b compare the performance of three feature selection algorithms when SIFT+FAP features are used. As seen be seen, mRMR and Adaboost achieve a similar performance and they outperform SVM with approximately 10% and 4% higher accuracy for the Feedtum and NVIE databases, respectively. Similar results are also observed when Gabor+FAP features are used. On the other hand, when LBP+FAP features are used, mRMR performs the best and the following one is SVM, while Adaboost ranks the last. Therefore, mRMR obtains the highest overall performance on both the databases, for all types of features and. Accordingly, the optimal combination is found to be SIFT+FAP+mRMR.

Figs. 4c and 5c demonstrate the performance obtained using three fusion strategies when SIFT and Adaboost are used. From the figures, we can observe that fusion of SIFT and FAP leads to higher accuracy on both the databases. In details, SIFT+FAP has approximately 3% and 14% higher accuracy than SIFT and FAP respectively on the Feedtum database, and 33% and 3% higher accuracy than SIFT and FAP on the NVIE database. Similar results in terms of performance improvement have also been observed for SIFT features when mRMR and SVM are used. When Gabor and LBP features are used, fused features also result in better performance on Feedtum for all selection algorithms. However, this is not always true for Gabor and LBP on NVIE, where Gabor+FAP+SVM and LBP+FAP+Adaboost lead to a little lower performance than FAP. This result is probably due to the sensitivity of LBP and Gabor features to big changes of the face size in NVIE images, where the changes are much larger than those in Feedtum images. Note that the comparative framework does not normalize the face size. On the contrary, FAP features are normalized using FAPUs and remain better insensitivity to these changes. It can be concluded that only SIFT benefits from fusing with FAP features for all three selection algorithms on the two databases. Fusing texture and FAP-based distances helps to improve the recognition performance for most of the cases.
Table 2 presents the best performance among all feature extraction and selection algorithms on the Feedtum and NVIE databases. Among all features, SIFT+FAP achieves the highest accuracy for all selection algorithms on both the databases. The highest accuracy 63.6% for Feedtum and 83.0% for NVIE are attained using SIFT+FAP and Adaboost (or mRMR for the case of NVIE). However, the performances between SIFT+FAP, LBP+FAP, and Gabor+FAP are not significantly different from each other on both the databases as shown by the one standard deviation. For all texture features, fusion with FAP features leads to 0.4-3.2% and 22-35.8% higher accuracy than using texture alone for Feedtum and NVIE respectively. Fused features also produce higher accuracy over FAP alone on Feedtum. In contrast, fusion may result in lower accuracy than FAP on NVIE. Three selection algorithms have similar performance for all features on Feedtum, and only slightly bigger differences in performance on NVIE. This indicates that the three selection algorithms can achieve a similar best performance based on the same feature set.

Higher performance of SIFT over LBP and Gabor can be explained as follows. SIFT and LBP descriptors represent the texture using histograms of orientation and histograms of binary values on pixel neighborhoods, respectively. The texture in the comparison framework is calculated around 53 key points, whose neighborhoods contain rich orientation information. Accordingly, SIFT captures the orientation information more effectively than LBP. The most discriminative features from LBP, on the other hand, may not all come from these key points, as shown by the LBP feature distribution in [10]. Gabor features are sensitive to shifting of key points from inaccurate ASM detection and big scale changes [3]. Our results also indicate that SIFT has a better tolerance to big changes of the face size than LBP and Gabor.

4.3. Computational time performance

Table 3 displays the average computational time per image used for calculating LBP, SIFT and Gabor features. The programs are developed in Matlab 7.6.0 under a laptop configuration of core duo 1.66GHz CUP and 2GB memory. Calculating SIFT demands about 35 and two times of the computational time for computing LBP and Gabor respectively. This is expected as SIFT contains computationally expensive steps (e.g. difference of Gaussian images). On the other hand, LBP requires the least time. NVIE images require more time to calculate all three features than Feedtum images, which is due to a larger size of the face in NVIE images than Feedtum images. For selection algorithms, mRMR requires much less processing time than Adaboost and SVM (time is not shown here). Thus, LBP and mRMR are good choices for developing systems demanding a fast speed.

4.4. Performance comparison with previous work

Table 4 compares the performance of the approach using SIFT+FAP and mRMR with previous work. Note that our approach does not require face normalization or depend on temporal information. When the Feedtum database is used, SIFT+FAP+mRMR obtains 1.3%, 29.4%, 20.7%, and 2.0% higher accuracy than the approaches using motion blocks, global motion, DCT features, and human perception in [22]. Wallhoff et al. [22] reported a mean recognition accuracy 61% when 20 subjects are asked to discriminate six basic plus neutral emotions. The method in [6] obtains higher classification accuracy, but it uses only a subset of images that have frontal faces without large head movements. In addition, it is based on only six emotions and also requires face registration. For the NVIE database, our method obtains 12.2% higher accuracy in classifying six emotions than that obtained classifying three emotions in [21]. Therefore, the approach using SIFT+FAP+mRMR achieves a state-of-the-art performance.
5. CONCLUSIONS

This paper proposes a framework for comparatively analyzing the performance using SIFT, LBP, Gabor and FAP-based distance features, and Adaboost, mRMR and SVM selection algorithms on recognizing spontaneous facial expressions. The experimental results on the Feedtum and NVIE databases show that fusion of texture and geometry leads to a higher recognition performance. SIFT+FAP+mRMR is the optimal combination for achieving high performance for spontaneous FER and also demonstrates a state-of-the-art performance. Both LBP and mRMR are good choices for achieving high accuracy and fast processing time.

6. REFERENCES