https://doi.org/10.2352/ISSN.2470-1173.2017.10.IMAWM-173

A Real-time Smile Elegance Detection System: A Feature-level Fusion and SVM Based Approach

Lili Lin[1], Yiwen Zhang[1], Weini Zhang[1], Zhihui Chen[1], Yan Yan[1], Tianli Yu[2]*
[1]School of Information Science and Engineering, Xiamen University, Fujian, P. R. China
[2]Morpx Inc., Mountain View, CA, U.S.A

Abstract

Smile detection in the unconstrained real-world scenario has attracted much attention due to its importance for mobile applications and human computer interaction. However, in many applications, how to determine the attractiveness of a smile face image (i.e., smile elegance) is an interesting task. In this paper, we present a real-time smile elegance detection system based on feature-level fusion and support vector machine (SVM). Specifically, three types of features, including local intensity histogram (LIH), central symmetry local binary pattern (CS-LBP) and Gabor wavelet, are firstly employed to characterize the global and local relationships of the face image. Then, a feature-level fusion method is leveraged to effectively combine these features. Finally, a specific SVM classifier is learnt to classify the smile face image into elegance or inelegance. Since there is no available public smile elegance database, we establish one for the first time. Experimental results on our collected smile elegance database demonstrate the effectiveness and efficiency of the proposed smile elegance detection system.

Introduction

As an important way to express emotion, facial expression has the vital influence on the communication between people. Currently, facial expression recognition has become an active research area in computer vision and pattern recognition. On one hand, as one of the most important expressions during the daily life communications, real-time and effective smile detection can significantly promote the development of facial expression recognition. Furthermore, smile detection is of great importance in practical applications, such as mobile applications and human computer interaction. On the other hand, based on the smile face image, how to further determine the attractiveness of a smile face image (we define it as **smile elegance** in this paper) becomes an interesting and important task. However, the related research on smile elegance has received little attention so far.

During the past few decades, smile detection has brought great interests both in academics and industry, and its related applications are widespread. For instance, the feature of automatic smile detection is used in a variety of digital cameras and mobile phones (e.g., the "OKAO Catch" system developed by the Omron Corporation). In general, current smile detection methods can be roughly divided into two categories: the feature-based methods [1-3], and the appearance-based methods [4-6]. In terms of the feature-based method, Nakano *et al.* [1] proposed to firstly use the sparse principal component analysis (SPCA) method to extract the features of eyes, nose and mouth. Then they combined these features with the nearest neighbor (NN) classifier to detect the smile face image. Ito *et al.* [2] employed the eigenvalues of the extracted feature points to constitute the feature vector for smile detection. Chen *et al.* [3] extracted the Haar-like features from the mouth areas, and then used the cascade classifier to classify the face image. For the appearance-based methods, Whitehill *et al.* [4] selected the Gabor energy and square wave as the features, and compared the performance of smile detection methods using different classifiers. Purnomo *et al.* [5] used the nonlinear mapping functions to deal with the image data, where the PCA transform and the Laplacian transform are respectively used. Yang *et al.* [6] proposed a novel face feature extraction algorithm based on the Gabor wavelet and the Adaboost algorithm, which effectively improves the speed and accuracy of smile detection.

Although the current studies on facial expression recognition, especially smile detection, have achieved remarkable results, most of the current researches focus on smile detection (or called smile intensity estimation). However, how to determine the attractiveness of a smile face (i.e., smile elegance) still needs further investigation. Note that smile elegance is totally different from the smile intensity. In this paper, we define the criterion of smile elegance as attractiveness and beautifulness determined subjectively. Figure 1 illustrates a pair of examples on elegant smile and inelegant smile. More specifically, in this paper, we propose a new research topic, i.e., **Smile Elegance Detection (SED)**, which extends the research of the smile detection to an emotional field and is closely related to the real-life applications.

Figure 1. An example of elegant smile and inelegant smile images.

In this paper, we present a real-time smile elegance detection system based on feature-level fusion and SVM. In particular, firstly, we apply face detection to locate the face image. Then, three types of features, including local intensity histogram (LIH), central symmetry local binary pattern (CS-LBP) and Gabor wavelet, are respectively employed to characterize the global and local relationships of the face image. At the same time, a feature-level fusion strategy is leveraged to effectively fuse these features. Finally, an SVM classifier is specifically trained for smile elegance detection. Experimental results on our collected elegance database (specifically established for this study), demonstrate the effectiveness and efficiency of the proposed smile elegance detection system.

* Corresponding Author

2. The Proposed SED System

In this section, the details of the proposed SED system are given as follows. In Section 2.1, face detection and preprocessing is shown. In Section 2.2, three types of features are respectively introduced, and the feature-level fusion method is also described. Finally, in Section 2.3, the SVM classifier for SED is introduced.

2.1 Face Detection and Preprocessing

Face detection is to determine the locations and sizes of all the faces in a static image or a video sequence, which is a key preprocessing step for SED. In this paper, we use the Adaboost face classifier with the Haar-like features to detect the faces based on OpenCV [7]. We briefly describe the Haar-like features, Adaboost algorithm, and cascade model used for face detection in the following.

The commonly used Haar-like features can be divided into four categories: edge features, line features, point features (center features), diagonal features. The value of the Harr-like feature is defined as the difference between the sums of the pixel intensities over two regions, which can be efficiently computed by using the integral image.

Given an image i, an integral image $ii(x, y)$ is defined as,

$$ii(x,y) = \sum_{x' \le x, y' \le y} i(x', y') \quad (1)$$

Here, the value of the integral image at the location (x, y) is the summation over all the left and upper pixel values of the original image i. Based on the integral image, the Haar-like features can be computed efficiently.

The Adaboost algorithm which is an iterative algorithm, is proposed by Freund *et al.* [8] on the basis of the Boosting algorithm. In the framework of Adaboost-based face detection, Adaboost is not only used to select discriminative features from the feature pool, but also can be considered as a classifier.

Finally, the cascade classifier model is used. Specifically, the cascade classifier is composed of several strong classifiers, each of which consists of a number of weak classifiers. The classifiers in the former stages are simple and can quickly filter out the background regions. The classifiers in the later stages are more complex so that they can spend more computational time on the promising face-like regions.

Based on the face detection results, the face can be cropped and aligned via the positions of eyes (note that the eye detector can be trained similar to the face detector). Then, each cropped face image is resized into the size of 40×40 pixels. The selection of ROI (Region of Interest) in a face image has an impact on the final performance of SED. We compare the following three different options: (a) the entire face; (b) the lower half of the face; (c) only the face part, as shown in Figure 2. In our experiments, the best performance is obtained when only the face part is used as the input, which demonstrates that the information in the whole face region is critical for SED.

(a) the entire face (b) lower half of the face (c) only the face part

Figure 2. Different ROIs of a face image.

2.2 Feature extraction

Feature extraction is a critical step for SED, which aims to build an informative and robust representation for the face. In this paper, we use the Local Intensity Histogram (LIH), Center-Symmetric Local Binary Pattern (CS-LBP) [9] and Gabor wavelet [10] features to analyze the smile elegance.

2.2.1 Local Intensity Histogram (LIH)

LIH [11] is generated by concatenating the intensity histograms of local regions. The specific feature extracting steps of LIH include: 1) divide the face image into P×Q cells; 2) build the intensity histogram with S bins for each cell; 3) normalize the intensity histogram of each cell; 4) concatenate the normalized histograms of all the cells to form a P×Q×S dimensional feature vector.

2.2.2 Center-Symmetric Local Binary Pattern (CS-LBP)

CS-LBP extracts the features similar to LBP, which encodes the image based on the central symmetry theory. CS-LBP generates a string of binary code by comparing the gray value of center symmetric pairs of pixels and then converts the binary code to the decimal code. CS-LBP is robust against illumination variations and face pose changes. Furthermore, it can effectively represent the information of the image texture. The CS-LBP is calculate by,

$$CS-LBP_{R,N,T}(x,y) = \sum_{i=0}^{(N/2)-1} s(n_i - n_{i+(N/2)})2^i,$$
$$s(x) = \begin{cases} 1, x > T \\ 0, otherwise \end{cases} \quad (2)$$

where T is an encoding threshold; N is the number of pixels on the circle of radius R centered at the center pixel; and n_i and $n_{i+(N/2)}$ correspond to the gray value of center symmetric pairs of pixels, respectively. In this paper, N is fixed to 8; R is fixed to 1; and T is fixed to 0. Similar to LIH, we can divide the face image into cells, and then build the CS-LBP histograms of all the cells to form a feature vector.

2.2.3 Gabor Wavelet

Gabor wavelet Transform is a kind of Windows Fourier Transform or Short time Fourier Transform [10]. When the windows function is defined as the Gaussian windows, it is commonly referred to the Gabor wavelet Transform. The filter function of Gabor wavelet Transform is similar to the stress reaction of human visual system [12]. Gabor filter can extract local spatial information of an object and can evaluate the object from different scales and orientations in the frequency domain. Gabor filter is sensitive to the edge of the image and it can enhance the local characteristics of parts that have effects on the facial expression, such as eyes, nose, mouth etc. The efficiency of the Gabor wavelet transform is largely determined by the efficiency of the convolution. The multi-scale and multi-resolution capability of Gabor filter in the spatial and frequency domains is mainly reflected in the kernel function with different frequencies and orientation parameters. The kernel function of the filter is usually defined as [13]:

$$G_k^t(\vec{z}) = \frac{|\vec{k}|^2}{\sigma^2} \exp\left(-\frac{|\vec{k}|^2}{2\sigma^2}|\vec{z}|^2\right)\left(\exp\left(i\vec{k} * \vec{z}\right) - \exp\left(-\frac{\sigma^2}{2}\right)\right) \quad (3)$$

where $\vec{z} = (x, y)$ is the position coordinate; $\vec{k}$ is the wavelet feature vector; and $\sigma / |\vec{k}|$ determines the size of Gaussian window.

The wavelet feature vector is then defined as:

$$\vec{k} = \begin{pmatrix} k_x \\ k_y \end{pmatrix} = \begin{pmatrix} k_y \cos\varphi_u \\ k_v \sin\varphi_u \end{pmatrix}, \quad k_v = 2^{-\frac{v+2}{2}}\pi, \quad \varphi_u = u\frac{\pi}{|\vec{k}|} \tag{4}$$

where v and u determines the wavelength and angle of Gabor filter, respectively; k_v is frequency; and φ_u is the orientation of filter.

Since the face image is usually a non-rigid body, the high frequency information of the face is more prominent. Therefore, we can select the kernel function with the high frequency, which contains more discriminative information for classification. In this paper, we select filter bank with 5 different scales and 8 different orientations [10]. Figure 3 shows 40 different filters. From left to right, the value of u is set to 0、π/8、2π/8、3π/8、4π/8、5π/8、6π/8、7π/8, respectively. From top to bottom, the value of v is set to 0、1、2、3、4, respectively.

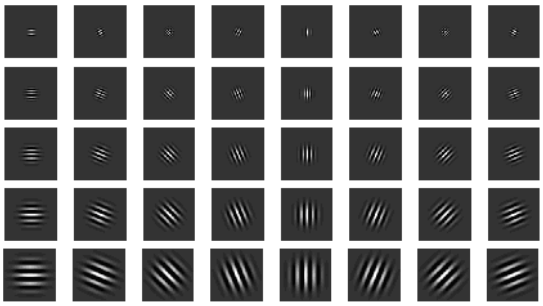

Figure 3. Gabor filter bank.

2.2.4 Feature-level Fusion

Based on the extracted features, we can fuse these features in different ways. In this paper, the feature-level fusion is employed, which not only can overcome the disadvantages of a single feature vector whose discriminability is limited, but also can enhance the performance of the final classifier.

In this paper we extract LIH, CS_LBP and Gabor features from the image. Then we combine the three features multiplied by three different scale factors in a cascade way. Specifically, based on the extracted LIH, CS_LBP, Gabor wavelet features, we form a feature vector by combining these three features as,

$$\vec{X} = \left(a \cdot \vec{X}_{LIH}, b \cdot \vec{X}_{CS-LBP}, c \cdot \vec{X}_{Gabor}\right) \tag{5}$$

where a, b, c are the scale factors; $\vec{X}_{LIH}$ represents the LIH features; $\vec{X}_{CS-LBP}$ denotes the CS_LBP features; $\vec{X}_{Gabor}$ represents the Gabor wavelet features.

2.3 Smile Elegance Detection

In this paper, based on the extracted features, support vector machine (SVM [14]) is used to perform smile elegance detection (SED). In machine learning, SVM is a popular supervised learning model with the associated learning algorithm which can effectively analyze the distribution of data. SVM has been widely used for classification and regression analysis. The main advantages of SVM can be summarized as: (1) it can be used for classification (both the linear separable and non-separable cases). For the linear non-separable case, the low-dimensional input space can be transformed to the high-dimensional feature space based on the kernel mapping. Therefore, it is possible to use the linear algorithm to analyze the non-linear distribution in the high-dimensional feature space. (2) Based on the structural risk minimization theory, the optimal partition hyperplane is constructed in the feature space, which makes SVM obtain the global optimized solution.

Different kernel functions can be used to generate different forms of SVM, and there are three different kinds of commonly used kernel functions, that is,

(1) Linear kernel function:

$$k(x_i, x_j) = x_i^T x_j \tag{6}$$

(2) Polynomial kernel function:

$$k(x_i, x_j) = [\gamma * (x_i^T x_j) + coef]^d \tag{7}$$

where d is the order of the polynomial and $coef$ is the offset coefficient.

(3) RBF kernel function:

$$k(x_i, x_j) = \exp(-\gamma * \mathrm{P}x_i - x_j\ \mathrm{P}^2) \tag{8}$$

where $\gamma > 0$ is the width of the kernel function.

The RBF kernel function performs the non-linear mapping from the low-dimensional space to the high-dimensional (or even infinite-dimensional) space, which can effectively characterize the nonlinear relationship between the class and the feature vector. The linear kernel function can be regarded as a special case of the RBF kernel function, and the polynomial kernel function is similar to the RBF kernel function under certain conditions. Therefore, the RBF kernel function has both the advantages of the linear kernel function and the polynomial kernel function. In addition, the RBF kernel function has only one adjustable parameter, which provides much convenience for parameter selection. According to the above discussions, we use SVM based on the RBF kernel function to perform SED. In order to obtain the optimal support vector machine parameter, 10-fold cross-validation is used in the RBF kernel function.

Based on the detected faces, we use the LIH, CS-LBP, Gabor wavelet features and fuse them in the feature-level. Then, we apply the SVM classifier to predict whether a smile face is elegant or not.

In summary, our proposed real-time smile elegance detection system mainly includes 3 stages: faced detection and pre-processing, feature extraction and fusion, and elegance detection. The whole framework of the proposed algorithm is summarized in Algorithm 1.

Algorithm 1 Real-time Smile Elegance Detection System

Input: An image

1. Detect face using the Adaboost face classifier based on the Haar-like features;
2. Crop and align the face image;
3. Perform Histogram equalization;
4. Extract LIH, CS_LBP, Gabor wavelet features;
5. Perform feature-level fusion;
6. Smile detection based on SVM;
7. If (face is smile)
   Perform smile elegance detection based on SVM;

Output: Smile elegance or smile inelegance

3. Experiments

In this section, we first introduce how to collect our smile elegance database in Section 3.1. Then the parameter settings of SVM are given in Section 3.2. The influence of different combinations is shown in Section 3.3. Finally, the discussions are presented in Section 3.4.

3.1 A New Database on Smile Elegance

To evaluate the performance of the proposed SED system, we need a standard database with positive and negative samples corresponding to smile elegance and smile inelegance, respectively. However, after reviewing several existing classical face databases, we find that there are no suitable databases for SED. Therefore, we collect a new database for smile elegance evaluation.

Figure 4. Elegant smile samples (1st and 2nd row) and inelegant smile samples (3rd and 4th row) belonging to the same identities.

The images of the database are manually collected from the Internet. We mainly choose the images from the celebrity. Initially, we select 500 samples of elegant smile, and 500 samples of inelegant smile, where each celebrity has a positive sample (i.e., elegant smile) and a negative sample (inelegant smile). As we know, the judgments of elegant and inelegant have strong subjectivity. Therefore, in order to improve the credibility of the database, we invite some volunteers (totally 10 persons) to determine whether the smile image is elegant or inelegant. The criterion of judgment for each sample is, if over 80% of the people believe the image is elegant, we label it as **elegant**. If over 50% of the people believe the image is inelegant, we label it as **inelegant**. Finally, we establish a smile elegance database consisting of 350 positive samples and 350 negative samples. Figure 4 shows some examples in the database.

3.2 Parameter Settings

Based on our collected database, we select 250 elegant smile images and 250 inelegant smile images for training. The remaining 100 elegant smile images and 100 inelegant smile images are used for performance evaluation. All the images are normalized to the size of 40×40 pixels. Three kinds of features, i.e., LIH, CS-LBP and Gabor wavelet features, are extracted and fused to characterize the smile elegance.

The two basic parameters of the SVM model are initially set as the following, i.e., the penalty factor $C = 10$, and the parameter of the kernel function $\gamma = 0.0001$. Then, we employ 10-fold cross-validation on C and γ to find the optimal values. The range of C is set as 2^{18}~2^{30} and the range of γ is set as 2^{-18}~2^{-10}. Experimental results (shown in Figure 5) show that when $C = 2^{20}$, $\gamma = 2^{-12}$, the correct rate of smile elegance detection achieves the highest. Therefore, we fix the values of C and γ to be 2^{20} and 2^{-12}, respectively in the following experiments.

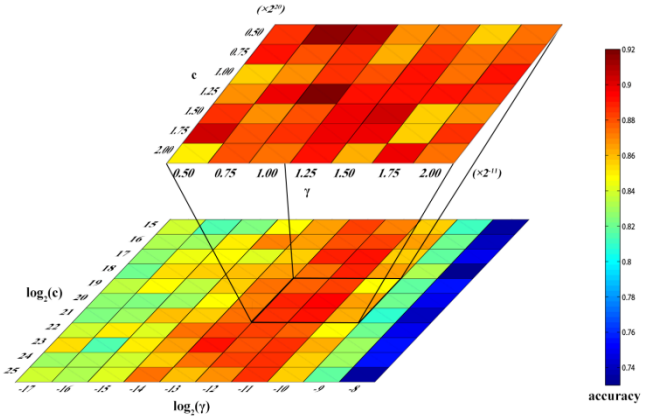

Figure 5. The results of the parameter optimization.

3.3 Influence of Different Features

Region of Interest (ROI) Selection We evaluate the selection of different ROIs to select the appropriate region for SED. We compare the following three choices: 1) the entire face image; 2) the lower half of the face image; 3) only the facial part of the face image (see Figure 2 for an illustration). The comparison results (i.e., the correct rates) based on different features are shown in Table 1.

Table 1. The correct rates of SED on different ROIs based on different features.

ROI	LIH	CS-LBP	Gabor
The entire face	71.0%	78.5%	75.0%
The lower half of the face	76.0%	80.5%	83.5%
Only the facial part	**87.5%**	**87.0%**	**90.0%**

As we can see from Table 1, when only the facial part is used as the ROI, the SED system achieves the top performance. The results prove that emphasizing on only the facial part can effectively avoid the distraction of hair, clothing and background. Besides, the regions around the eyes and eyebrows are also critical for SED. Therefore, in the following experiments, we set the ROI to be only the facial part.

In Table 1, we demonstrate the correct rate of SED system based on the single features (i.e., LIH, CS-LBP, Gabor respectively). Next, we evaluate the performance of the SED system based on different feature-level fusions in the following

experiments. The results are shown from Figure 6 to Figure 9. In these figures, the abscissa represents different weighted fusions of the features. The left coordinate represents the number of correctly recognized images. The red band represents the number of correctly detected elegant smile images while the blue band represents the number of correctly detected inelegant smile images. The right coordinate indicates the correct rate of the total smile images, which is indicated by the light green line in the figures.

LIH and CS-LBP Features The combination results of LIH and CS-LBP are shown in Figure 6. When the ratio between the weights of LIH and CS-LBP is 1:1, the detection accuracy achieves the highest (up to 90.0%).

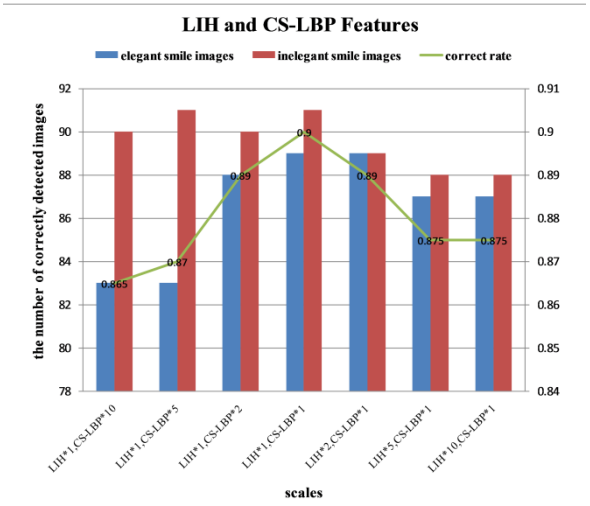

Figure 6. Smile elegance detection results on different weighted combinations of LIH and CS-LBP features.

LIH and Gabor Features The combination results of LIH and Gabor are shown in Figure 7. When the ratio between the weights of LIH and Gabor is 5:1, the detection accuracy achieves the highest (up to 90.0%).

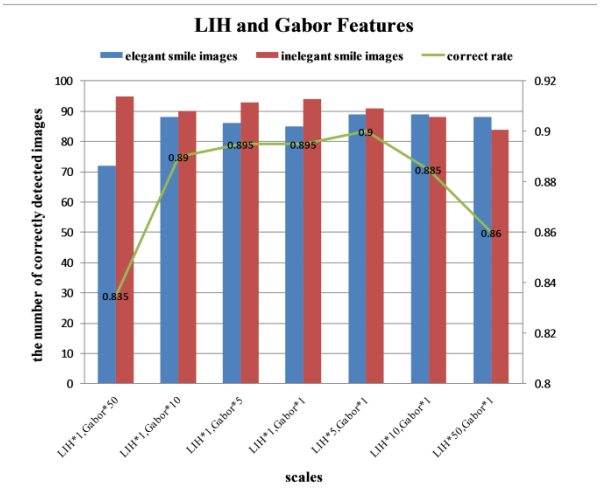

Figure 7. Smile elegance detection results on different weighted combinations of LIH and Gabor features.

CS-LBP and Gabor Features The combination results of CS-LBP and Gabor features are shown in Figure 8. When the ratio between the weights of CS-LBP and Gabor is 10:1, the detection accuracy achieves the highest (up to 91.0%).

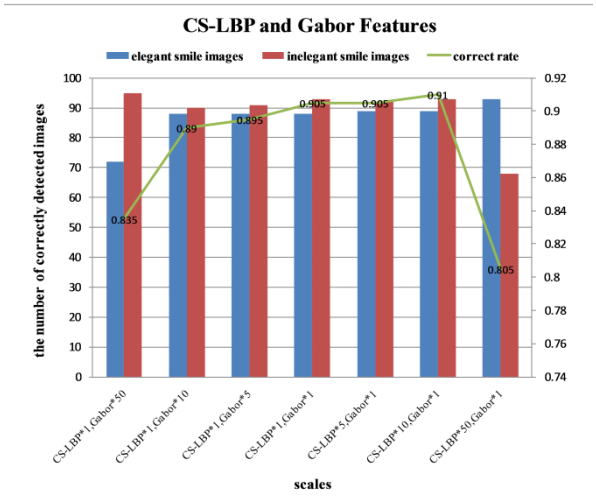

Figure 8. Smile elegance detection results on different weighted combinations of CS-LBP and Gabor features.

LIH, CS-LBP and Gabor Features The combination results of LIH, CS-LBP and Gabor are shown in Figure 9. When the ratio between the weights of LIH, CS-LBP and Gabor is 1:10:1, the detection accuracy achieves the highest, up to 91.5%.

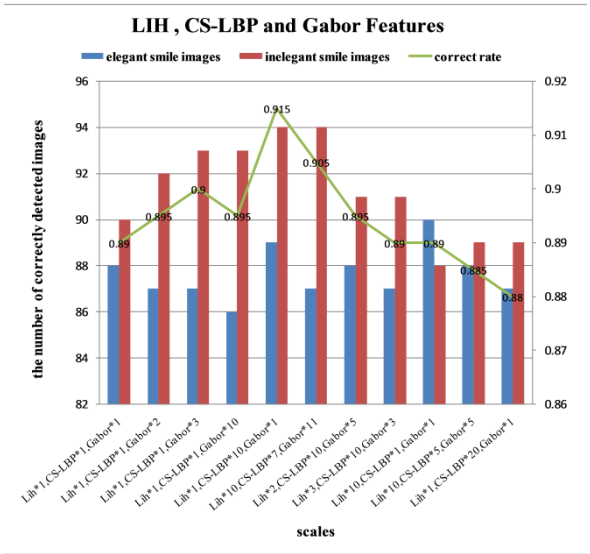

Figure 9. Smile elegance detection results on different weighted combinations of LIH, CS-LBP and Gabor features.

3.4 Discussions

Figure 10 shows some representative examples in correct classification and error classification of positive and negative samples. We can summarize the main reasons for error classify-cation as following, such as the face facing to the left direction or right direction (e.g., e37), the chin covered by hands (e.g., e13), facial features are not clear because of strong lighting conditions (e.g., i36) and the mouth location is not correctly located (e.g., i52).

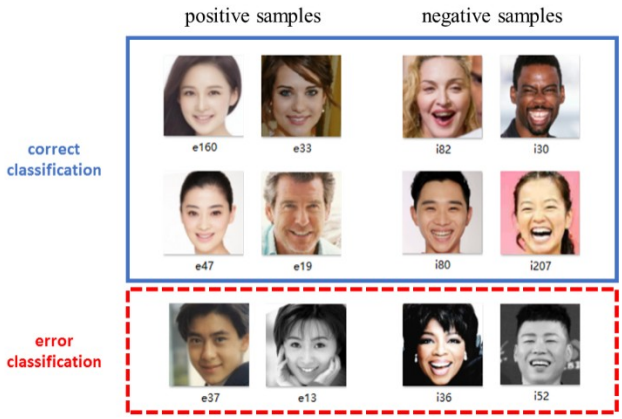

Figure 10. Positive smile samples (1st and 2nd column) and negative smile samples (3rd and 4th column) and correct classification (1st and 2nd row) and error classification (3rd row).

4. Conclusion

In this paper, we present a real-time smile elegance detection system based on feature-level fusion and support vector machine (SVM). Firstly, the detected face is cropped and aligned based on the eye positions, and the size of the image is normalized. Secondly, three types of features, including local intensity histogram (LIH), central symmetry local binary pattern (CS-LBP) and Gabor wavelet, are employed to perform feature extraction. Thirdly, a feature-level fusion strategy is leveraged to effectively combine these features. Finally, a robust smile classifier based on SVM is trained to build the real-time smile elegance detection system. As we can see from the experimental results, the proposed smile elegance detection system has shown the promising performance on our collected database.

5. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 61571379, and 61503315, and also supported by the Fundamental Research Funds for the Central Universities under Grants 20720162012.

References

[1] M. Nakano, Y. Mitsukura, M. Fukumi et al. True smile recognition system using neural networks[C]. IEEE Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'02), 2:650-654, 2002.

[2] A. Ito, X. Wang, M. Suzuki, S. Makino. Smile and laughter recognition using speech processing and face recognition from conversation video[C]. Proceedings of the 2005 IEEE International Conference on Cyberworlds, 8:437-444, 2005.

[3] J. Y. Chen, O. Lemon. Facial feature detection and tracking in a new multimodal technology-enhanced learning environment for social communication[C]. IEEE International Conference on Signal and Image Processing Applications, 279-284, 2009.

[4] J. Whitehill, G. Littlewort, I. Fasel et al. Toward practical smile detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(11):2106-2111, 2009.

[5] M. H. Purnomo, T. A. Sarjono, A. Muntasa. Smile stages classification based on kernel Laplacian-lips using selection of non linear function maximum value[C]. IEEE International Conference on Virtual Environments Human-Computer Interfaces and Measurement Systems (VECIMS), 151-156, 2010.

[6] W. G. Yang, S. H. Zheng. A fast smile recognition method for mobile phone platform[J]. Journal of System Simulation, 24(1), 2012.

[7] P. Viola, M. Jones. Rapid object detection using a boosted cascade of simple features[C]. 2001. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR01), 1(1):511-518, 2001.

[8] Y. Freund, R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting[A]. Second European Conference on Computational Learning Theory (EuroCOLT'95), 23-37, 1995.

[9] M. Heikkilä, M. Pietikäinen, C. Schmid. Description of interest regions with local binary patterns[J]. Pattern Recognition, 42(3):425-436, 2009.

[10] T. S. Lee. Image representation using 2D Gabor wavelets[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(10):959-971, 1996.

[11] K. Shimada, Y. Noguchi, T. Kuria. Fast and robust smile intensity estimation by cascaded support vector machines[J]. International Journal of Computer Theory & Engineering, 5(1), 2013.

[12] S. Franco. Design with Operational Amplifiers and Analog Integrated Circuits[M]. McGraw-Hill, 2002.

[13] J. G. Daugman. Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 36(7):1169-1179, 1988.

[14] C. Cortes, V. Vapnik. Support-vector networks. Machine Learning, 20(3):273–297, 1995.