1-1-2010

Improved facial expression recognition with trainable 2-D filters and support vector machines

Peiyao Li
University of Wollongong, pli@uow.edu.au

Son Lam Phung
University of Wollongong, phung@uow.edu.au

Abdesselam Bouzerdoum
University of Wollongong, bouzer@uow.edu.au

Fok Hing Chi Tivive
University of Wollongong, tivive@uow.edu.au

Recommended Citation

Li, Peiyao; Phung, Son Lam; Bouzerdoum, Abdesselam; and Tivive, Fok Hing Chi: Improved facial expression recognition with trainable 2-D filters and support vector machines 2010, 3732-3735.
https://ro.uow.edu.au/infopapers/815

Improved facial expression recognition with trainable 2-D filters and support vector machines

Abstract

Facial expression is one way humans convey their emotional states. Accurate recognition of facial expressions is essential in perceptual human-computer interface, robotics and mimetic games. This paper presents a novel approach to facial expression recognition from static images that combines fixed and adaptive 2-D filters in a hierarchical structure. The fixed filters are used to extract primitive features. They are followed by the adaptive filters that are trained to extract more complex facial features. Both types of filters are non-linear and are based on the biological mechanism of shunting inhibition. The features are finally classified by a support vector machine. The proposed approach is evaluated on the JAFFE database with seven types of facial expressions: anger, disgust, fear, happiness, neutral, sadness and surprise. It achieves a classification rate of 96.7%, which compares favorably with several existing techniques for facial expression recognition tested on the same database.

Keywords

filters, 2, machines, trainable, recognition, expression, facial, vector, improved, support

Publication Details

Li, P., Phung, S., Bouzerdoum, A. & Tivive, F. (2010). Improved facial expression recognition with trainable 2-D filters and support vector machines. Proceedings of the 20th International Conference on Pattern Recognition (ICPR 2010) (pp. 3732-3735). USA: IEEE.

Improved Facial Expression Recognition with Trainable 2-D Filters and Support Vector Machines

P. Li, S. L. Phung, A. Bouzerdom, and F. H. C. Tivive
School of Electrical, Computer and Telecommunication Engineering,
University of Wollongong, Wollongong, NSW 2522, Australia

Abstract

Facial expression is one way humans convey their emotional states. Accurate recognition of facial expressions is essential in perceptual human-computer interface, robotics and mimetic games. This paper presents a novel approach to facial expression recognition from static images that combines fixed and adaptive 2-D filters in a hierarchical structure. The fixed filters are used to extract primitive features. They are followed by the adaptive filters that are trained to extract more complex facial features. Both types of filters are non-linear and are based on the biological mechanism of shunting inhibition. The features are finally classified by a support vector machine. The proposed approach is evaluated on the JAFFE database with seven types of facial expressions: anger, disgust, fear, happiness, neutral, sadness and surprise. It achieves a classification rate of 96.7%*, which compares favorably with several existing techniques for facial expression recognition tested on the same database.*

1. Introduction

Facial expressions play a vital role in human communication, enabling us to convey emotions almost instantly. Humans are capable of producing up to $7,000$ different facial expressions and reading other people's expressions accurately. Facial expression recognition (FER) aims to automatically determine facial expressions from images or video sequences. It is a vision task with many applications in perceptual human-computer interaction, robotics, computer games, and psychology studies.

Based on the type of features, existing approaches to facial expression recognition can be divided into two categories: geometric-based and appearance-based. In geometric-based approaches, a face image is usually represented geometrically via fiducial points [4] or the shape of facial regions [8]. Classification is done by analyzing the distances between feature points and the relative sizes of the facial components. Pantic *et al*. [8] proposed a multi-detector approach to analyze the spatial changes in the contour of facial components such as the eyes. Based on this information, a rule-based method was used to describe the facial actions. Geometric-based methods can cope well with variations in skin patterns or dermatoglyphics. However, they usually require accurate detection of facial fiducial points, which is difficult when the image has a low-quality or a complex background.

In appearance-based approaches, a face image is processed as a whole. These approaches typically use image filters to extract facial features from the entire face or a specific region. Feng [3] used Local Binary Patterns (LBP) to extract facial texture features and combined different local histograms to recover the shape of the face. A coarse-to-fine classification scheme is utilized to detect appearance changes and differentiate the facial expressions. Zhen *et al*. [10] used Gabor wavelets to extract appearance changes as a set of multi-scale and multi-orientation coefficients. Experimental results show this approach can cope with different people and illumination conditions.

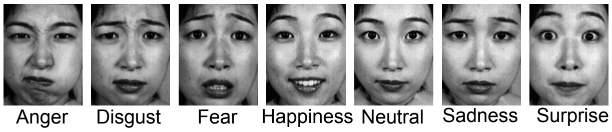

Figure 1. Examples of facial expressions.

In this paper, we propose a new approach to recognize facial expressions from static images using appearance features. In this approach, fixed and adaptive nonlinear 2-D filters are combined in a hierarchical structure. The fixed filters are used to extract primitive features such as edges and orientations, whereas the adaptive filters are trained to extract more complex and subtle facial features for classification. Here, we focus

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.909

on seven basic facial expressions that reflect distinctive psychological activities: anger, disgust, fear, happiness, neutral, sadness and surprise. Examples of these facial expressions are shown in Figure 1.

The paper is organized as follows: Section 2 presents the proposed method. Section 3 analyzes the performance of the proposed method on a standard database, and compares it with several existing techniques. Section 4 gives concluding remarks.

2. Proposed method

The proposed system consists of three processing stages as shown in Figure 2. The first and second stages consist of nonlinear filters, which are used for extracting 2-D visual features. The third stage performs classification.

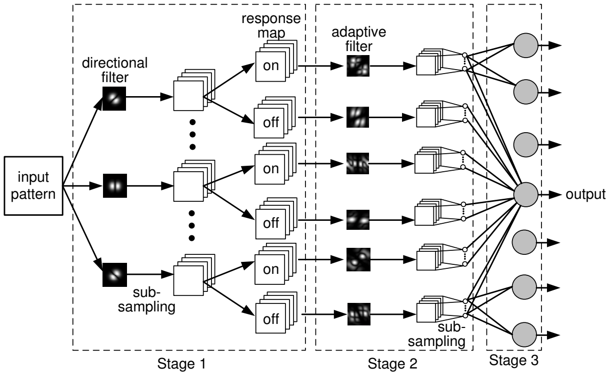

Figure 2. Block diagram of the proposed system.

2.1 Stage 1 - Directional Filters

Stage 1 is designed to extract features at different orientations. It consists of a set of nonlinear filters that are based on a biological mechanism known as the *shunting inhibition*. This mechanism, found in the cortical cells of the human visual system, has been applied to improve image contrast [5]. The output of the proposed directional nonlinear filter is computed as

$$\mathbf{Z}_{1,i} = \frac{\mathbf{D}_i * \mathbf{I}}{\mathbf{G} * \mathbf{I}}, \tag{1}$$

where $\mathbf{I}$ is a 2-D input face pattern, $\mathbf{Z}_{1,i}$ is the output of the i-th filter, $\mathbf{D}_i$ and $\mathbf{G}$ are the filter coefficients, and "$*$" denotes 2-D convolution. In this paper, the subscripts 1 and 2 in $\mathbf{Z}_{1,i}$ and $\mathbf{Z}_{2,i}$ indicate the outputs of the first and second processing steps, respectively. The kernel $\mathbf{G}$ is chosen as an isotropic Gaussian kernel:

$$\mathbf{G}(x,y) \quad = \quad \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}). \tag{2}$$

To extract elementary facial features at different directions, the kernel $\mathbf{D}_i$ is formulated as the M-th order derivative Gaussian. Its coefficients is defined as

$$\mathbf{D}_i(x,y) = \sum_{k=0}^{M} \frac{M!}{k!(M-k)!} s_x^k s_y^{M-k} \frac{\partial^M \mathbf{G}(x,y)}{\partial x^k \partial y^{M-k}}, \tag{3}$$

where θ_i is the angle of rotation, $s_x = \sin\theta_i$ and $s_y = \cos\theta_i$. Note that, $\theta_i = (i-1)\pi/N_1$ for $i = 1, 2, ..., N_1$.

The partial derivative of the Gaussian with respect to image coordinate x or y can be computed as the product of the Hermite polynomial and the Gaussian function,

$$\frac{\partial^k \mathbf{G}(x,y)}{\partial x^k} = \frac{(-1)^k}{(\sqrt{2}\sigma)^k} H_k(\frac{x}{\sqrt{2}\sigma})\mathbf{G}(x,y), \tag{4}$$

where $H_k(.)$ is the Hermite polynomial of order k.

Robust image classification requires visual features that are tolerant to small translations or geometric distortions in the input image. To achieve this, we perform a sub-sampling operation and decompose each filter output $\mathbf{Z}_{1,i}$ into four smaller maps, as shown in Figure 3a:

$$\mathbf{Z}_{1,i} \rightarrow \{\mathbf{Z}_{2,4i-3}, \mathbf{Z}_{2,4i-2}, \mathbf{Z}_{2,4i-1}, \mathbf{Z}_{2,4i}\}. \tag{5}$$

The first map $\mathbf{Z}_{2,4i-3}$ is formed from the odd rows and odd columns in $\mathbf{Z}_{1,i}$; the second map $\mathbf{Z}_{2,4i-2}$ is formed from the odd rows and even columns, and so on.

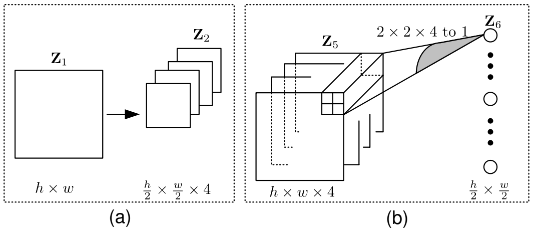
$\mathbf{Z}_1$
$\mathbf{Z}_2$
$h \times w$
$\frac{h}{2} \times \frac{w}{2} \times 4$
(a)
$\mathbf{Z}_5$
$2 \times 2 \times 4$ to 1
$\mathbf{Z}_6$
$h \times w \times 4$
$\frac{h}{2} \times \frac{w}{2}$
(b)

Figure 3. The sub-sampling operations performed in (a) Stage 1 and (b) Stage 2.

The next processing step is motivated by the center-surround receptive fields and the two configurations on-center and off-center in the human visual system. We separate each sub-sampled map $\mathbf{Z}_{2,i}$, where $i = 1, 2, ..., 4N_1$, into an on-response map and an off-response map, using zero as a threshold:

$$\mathbf{Z}_{2,i} \rightarrow \begin{cases} \text{on} \ : \ \mathbf{Z}_{3,2i-1} = \max(\mathbf{Z}_{2,i}, 0) \\ \text{off} \ : \ \mathbf{Z}_{3,2i} = -\min(\mathbf{Z}_{2,i}, 0) \end{cases} . \tag{6}$$

Essentially, for the on-response map, all negative entries are set to 0, whereas for the off-response map, positive entries are set to 0 and the entire map is then negated.

Next, each map is contrast-normalized using the transformation equation: $\mathbf{Z}_{4,i} = \mathbf{Z}_{3,i}/(\mathbf{Z}_{3,i} + \mu_i)$, where μ_i is the mean value of the map.

2.2 Stage 2 - Trainable Filters

Stage 2 aims to detect more complex features for classification. The output maps produced by each filter in Stage 1 are processed by exactly two filters in Stage 2: one filter for the on-response and the other filter for the off-response. Hence, the number of filters, N_2, in Stage 2 is twice the number of filters in Stage 1: $N_2 = 2N_1$.

Stage 2 is also based on the shunting inhibition mechanism. Consider an input map $\mathbf{Z}_{4,i}$ to Stage 2. Suppose that $\mathbf{P}_k$ and $\mathbf{Q}_k$ are two adaptive convolution masks for the filter that corresponds to this input map. The filter output is calculated as

$$\mathbf{Z}_{5,i} = \frac{g\Big(\mathbf{P}_k * \mathbf{Z}_{4,i} + b_k\Big) + c_k}{a_k + f\Big(\mathbf{Q}_k * \mathbf{Z}_{4,i} + d_k\Big)}, \tag{7}$$

where a_k, b_k, c_k and d_k are adjustable bias terms, and f and g are two activation functions.

To form a feature vector, a sub-sampling operation is performed across each set of four output maps. From four output maps, each non-overlapping block of size $(2 \times 2$ pixels$) \times (4$ maps$)$ is averaged into a single output signal, as shown in Figure 3b:

$$\{\mathbf{Z}_{5,4i-3}, \mathbf{Z}_{5,4i-2}, \mathbf{Z}_{5,4i-1}, \mathbf{Z}_{5,4i}\}. \tag{8}$$

2.3 Stage 3 - Classification

The extracted features are sent to Stage 3 for classification. Stage 3 may use any type of classifier. To train the 2-D adaptive filters in Stage 2, we first use a simple linear classifier whose output y_j is given as

$$y_j = \sum_{i=1}^{N_3} w_{ij}\, \mathbf{Z}_{6,i} + b_j, \qquad j = 1, 2, ..., N_4 \tag{9}$$

where w_{ij}'s are adjustable weights, b_j is an adjustable bias term, $\mathbf{Z}_{6,i}$'s are input features to Stage 3, N_3 is the number of input features, and N_4 is the number of output nodes. The output $\mathbf{y} = [y_1, y_2, ..., y_{N_4}]^T$ indicates the class or the label of the input pattern $\mathbf{I}$.

After the 2-D adaptive filters are found for Stage 2, we train and use support vector machine (SVM) classifier for Stage 3. This strategy is adopted to improve classification performance. SVM is an important tool in pattern classification. It has been shown to achieve good generalization by maximizing the margin between two classes. SVM has been developed initially for two-class problems. To solve multi-class problems, we can construct several SVMs to differentiate each pair of classes. For example, for seven facial expressions, we need 21 pair-wise SVMs.

To implement this training approach, we adopt a two-step process. In Step 1, we assume that a linear classifier is used in Stage 3, and calculate the coefficients of filters in Stage 2 and the weights of the linear classifier via supervised learning, using the Levenberg-Marquardt (LM) and least-square methods. In Step 2, once the filters in Stage 2 are found, we train the multi-class SVM with the RBF kernels to classify the extracted features.

3. Results and analysis

In this section, we analyze the performance of the proposed method on a benchmark facial expression data set. We also compare the proposed method and other existing methods for facial expression recognition.

3.1 Database and experimental steps

The proposed system is evaluated on the Japanese Female Facial Expression (JAFFE) database [7], which is commonly used in research on facial expression recognition. This database consists of 213 images from 10 Japanese actresses. They were instructed to produce seven types of facial expressions (see Figure 1). For each person, two to four images were recorded for each facial expression.

We apply the 10-fold cross validation on the JAFFE database, as in [4]. All images are divided into ten groups. For each validation fold, nine groups of them are used to train the classifier while the remaining group is used for testing. This step is repeated five times, and the classification rates of the ten folds are averaged to form the final estimate of the classification rate.

The proposed system uses an input image size of 44×32 pixels. Stage 1 uses the second-order Gaussian derivative ($M = 2$) and four directions ($N_1 = 4$). The filter sizes for Stages 1 and 2 are 7-by-7 and 3-by-3 pixels, respectively. Our experiments use the LIB-SVM package, developed by Chang and Lin at National Taiwan University [2]. To improve accuracy, we present both the input pattern and its mirror image to the classifier, and then use the averaged response to form a classification decision.

3.2 Classification accuracy

The classification rates for different facial expressions are shown in Table 1. The entry (at row r, column c)

Table 1. Classification rates for different facial expression categories. Method: hybrid filters + SVM + mirror image.

%	AN	DI	FE	HA	NE	SA	SU
Anger	**96.7**	0.0	0.0	0.0	0.0	3.3	0.0
Disgust	0.0	**96.6**	0.0	0.0	0.0	3.4	0.0
Fear	0.0	0.0	**93.7**	0.0	0.0	0.0	6.3
Happiness	0.0	0.0	0.0	**100**	0.0	0.0	0.0
Neutral	0.0	0.0	0.0	0.0	**100**	0.0	0.0
Sadness	0.0	0.0	0.0	3.2	0.0	**96.8**	0.0
Surprise	0.0	0.0	0.0	6.7	0.0	0.0	**93.3**

Table 2. Classification rates of FER methods on JAFFE database.

Method	**CR** (%)
Hybrid filters + SVM (RBF) + mirror	96.7
Hybrid filters + SVM (RBF)	96.2
Hybrid filters + Linear classifier + mirror	95.9
Hybrid filters + Linear classifier	95.3
Gabor + Linear SVM [1]	95.2
Fiducial points + FSLP [4]	91.0
Gabor + MLP [6]	90.2
Fiducial points + two-layer MLP [9]	90.1
LBP + Coarse-to-fine [3]	77.0
Fiducial points + AdaBoost [4]	71.9
Fiducial points + Bayes rule [4]	71.0

is the percentage of facial expression r that is classified as facial expression c. For example, 96.7% of anger expressions are correctly classified as anger, whereas 3.3% of anger expression are misclassified as sadness.

The classification rates for the seven facial expressions are: anger 96.7%, disgust 96.6%, fear 93.7%, happiness 100.0%, neutral 100.0%, sadness 96.8% and surprise 93.3%. The system can recognize happiness and neutral expressions well. It can recognize anger, disgust, and sadness expressions better than fear and surprise expressions.

Table 2 shows the classification rates of several FER methods, tested on the JAFFE database using ten-fold validation. Guo and Dyer [4] compared several feature selection schemes: all features, feature selection via linear programming (FSLP), and feature selection via adaptive boosting (AdaBoost). Busiu *et al*. [1] used Gabor wavelets to extract image features and linear SVM as a classifier. Zhang *et al*. [9] used 34 manually defined fiducial points in feature extraction. Koutlas and Fotiadis [6] used 20 automatically defined fiducial points and feed-forward neural networks (MLP). The proposed system with and without mirror image has classification rates of 96.7% and 96.2%, respectively. It has higher classification rates compared to existing FER methods.

4. Conclusion

We presented an approach for facial expression recognition that is based on fixed, directional filters and adaptive filters connected in a hierarchical structure. The directional filters extract primitive facial features, whereas adaptive filters are trained to extract more complex features, which are then classified by an SVM. The proposed system has a classification rate of 96.2%, which is higher than existing methods tested on the JAFFE database. Furthermore, by combining several SVMs and using the mirror image, the classification rate is improved to 96.7%. For future research, we plan to train the SVM in Stage 3 simultaneously with the adaptive filters in Stage 2.

References

[1] I. Buciu, C. Kotropoulos, and I. Pitas. ICA and Gabor representation for facial expression recognition. In *Proc. ICIP*, pages 855–858, 2003.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for Support Vector Machines, 2001.

[3] X. Feng. Facial expression recognition based on local binary patterns and coarse-to-fine classification. In *Proc. ICCIT*, pages 178–183, 2004.

[4] G. Guo and C. R. Dyer. Learning from examples in the small sample case: face expression recognition. *IEEE Trans. SMC, Part B: Cybernetics*, 35(3):477–488, 2005.

[5] T. Hammadou and A. Bouzerdoum. Novel image enhancement technique using shunting inhibitory cellular neural networks. *IEEE Trans. Consumer Electronics*, 47(4):934–940, 2001.

[6] A. Koutlas and D. I. Fotiadis. An automatic region based methodology for facial expression recognition. In *Proc. IEEE SMC*, pages 662–666, 2008.

[7] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proc. IEEE FG*, pages 200–205, 1998.

[8] M. Pantic and L. J. M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Trans. SMC, Part B: Cybernetics*, 34(3):1449–1461, 2004.

[9] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Proc. IEEE FG*, pages 454–459, 1998.

[10] W. Zhen and T. S. Huang. Capturing subtle facial motions in 3D face tracking. In *Proc. ICCV*, pages 1343–1350, 2003.