TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 15/15 pp216-224
Volume 16, Number 2, April 2011

Histogram of the Oriented Gradient for Face Recognition*

SHU Chang (舒 畅), DING Xiaoqing (丁晓青)**, FANG Chi (方 驰)

State Key Laboratory of Intelligent Technology and System, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

Abstract: The histogram of oriented gradient has been successfully applied in many research fields with excellent performance especially in pedestrian detection. However, the method has rarely been applied to face recognition. Aimed to develop a fast and efficient new feature for face recognition, the original HOG and its variations were applied to evaluate the effects of different factors. An information theory-based criterion was also developed to evaluate the potential classification power of different features. Comparative experiments show that even with a relatively simple feature descriptor, the proposed HOG feature achieves almost the same recognition rate with much lower computational time than the widely used Gabor feature on the FRGC and CAS-PEAL databases.

Key words: face recognition; feature; histogram of oriented gradient

Introduction

The intensity of an image contains discriminative information as well as noise, and in most cases, is the only source that can be used to still object recognition. However, what really matters is not the absolute value, but the relative value which reflects the structure information or texture variation of an object.

Various feature extraction and selection methods have been widely used[1-5]. Besides holistic methods such as PCA and LDA, local descriptors have been studied recently. An ideal descriptor for the local facial regions should have large inter-class variance and small intra-class variance, which means that the descriptor should be robust with respect to varying illumination, slight deformations, image quality degradation, and so on. Information theory was used to develop a criterion to evaluate the potential classification power of different features.

Among the variety of different descriptors for the appearance of image patches that have been developed by the texture analysis community, local binary pattern features yield some of the best results when used to represent facial images. The idea of using a local binary pattern for facial descriptions is that faces can be seen as a composite of micro-patterns which are well described by this operator[6]. However, sometimes there are too many micro patterns so in practice a system has to reduce the number of local regions or the number of possible scales to form a reasonable length feature vector.

The Gabor wavelet[7], whose kernels are similar to the 2-D representative profiles of the mammalian cortical simple cells, was first introduced by Gabor in 1946. The Gabor transformation simultaneously enhances facial feature magnitude and orientation and has been widely used as an effective element in image processing and pattern recognition tasks. The Gabor wavelet is the most popular and successful feature ever used for face recognition. For example, it is used for face recognition in the dynamic link architecture

Received: 2010-02-10; revised: 2010-10-19

* Supported by the National Key Basic Research and Development (973) Program of China (No. 2007CB311004) and the National High-Tech Research and Development (863) Program of China (No. 2006AA01Z115)

** To whom correspondence should be addressed.
E-mail: dingxq@tsinghua.edu.cn; Tel: 86-10-62772368

framework by Lades et al.[8] and in the elastic bunch graph matching method developed by Wiskott et al.[9]

The use of orientation histograms also has many precursors. Freeman and Roth[10] used orientation histograms for hand gesture recognition, Dalal and Triggs[11] presented a pedestrian detection algorithm with excellent detection results using a dense grid of HOG. The HOG provides the underlying image patch descriptor for matching scale invariant keypoints when combined with local spatial histogramming and normalization in Lowe's scale invariant feature transformation (SIFT) approach to wide baseline image matching[12]. However, few publications can be found which show a successful application of this feature to face recognition.

1 Histogram of Oriented Gradient

1.1 Basic theory

The basic idea of HOG features is that the local object appearance and shape can often be characterized rather well by the distribution of the local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. The orientation analysis is robust to lighting changes since the histogramming gives translational invariance. The HOG feature summarizes the distribution of measurements within the image regions and is particularly useful for recognition of textured objects with deformable shapes. The method is also simple and fast so the histogram can be calculated quickly.

As used in SIFT or the EBGM method, the original HOG feature is generated for each key point of an image. The neighboring area around each key point is divided into several uniformly spaced cells and for each cell a local 1-D histogram of gradient directions or edge orientations is accumulated over all the pixels of the cell. The histogram entries of all cells around one key point form the feature of that key point. The combined histogram features of all key points form the image representation. The whole process is shown in Fig. 1.

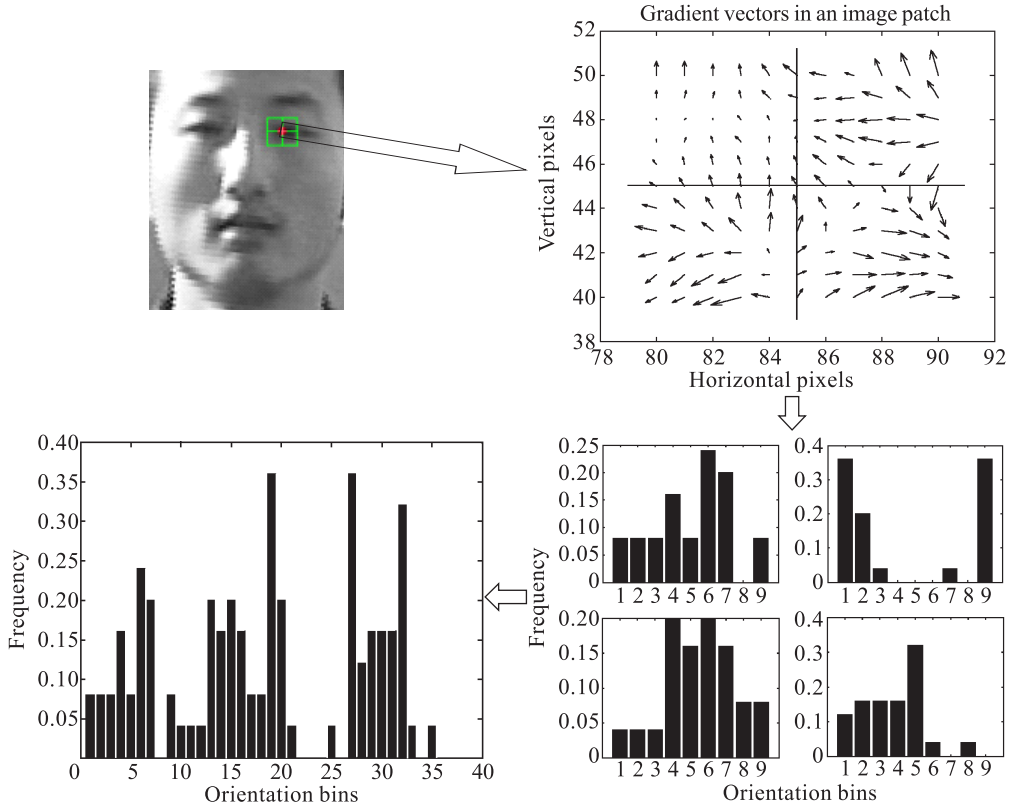


Fig. 1 Image window divided into small spatial regions ("cells"). Local 1-D histograms of gradient directions or edge orientations are accumulated and concatenated to form the final histogram feature.

1.2 Orientation representation

Orientation can be represented as a single angle or as a double angle[13]. A single angle treats a given edge and a contrast reversed region as having opposite orientations. A double angle representation maps these into

the same orientation. The single angle representation may allow more patterns to be distinguished. This work used a single angle representation to allow more differentiation between patterns. Tests in part 4 show that the single angle representation performs much better than the double angle representation. Note that this differs from the classic Gabor feature, which uses a single angle representation instead of the double angle.

If an image window I of a key point is uniformly divided into N cells, the image window can be represented as

$$I = \bigcup_{t=1}^{N} C_t \tag{1}$$

where C_t is the set of all pixels belonging to the t-th cell. For any pixel $p(x,y)$ of the image window I, the contrast is given by

$$g_p = g(x,y) = \sqrt{\Delta x^2 + \Delta y^2} \tag{2}$$

and the gradient direction is given by

$$\theta_p = \theta(x,y) = \arctan\frac{\Delta y}{\Delta x} \tag{3}$$

If the orientation is divided into H bins, which means the histogram vector length for each cell is H, the histogram vector can be calculated as follows:

$$b_t^i = \frac{\sum\{g_p \mid p \in C_t, \theta_p \in [i\theta_0 - \theta_0/2, i\theta_0 + \theta_0/2]\}}{|C_t|} \tag{4}$$

$$\boldsymbol{v}_t = \{b_t^0, b_t^1, b_t^2, \cdots, b_t^{H-1}\} \tag{5}$$

$|C_t|$ denotes the size of set C_t.

1.3 Normalization

For better invariance to illumination and noise, a normalization step is usually used after calculating the histogram vectors. Four different normalization schemes have been proposed[11]: L2-norm, L2-Hys, L1-sqrt, and L1-norm. This analysis used the L2-norm scheme due to its better performance:

$$\boldsymbol{v}_t' = \boldsymbol{v}_t / \sqrt{\|\boldsymbol{v}_t\|_2^2 + \varepsilon^2} \tag{6}$$

where ε is a small positive value used for some regularization when an empty cell is taken into account.

1.4 Fast computation

Liu et al.[14] introduced methods for fast computations of histogram bin weights for pixels whose gradient orientations are not in the orientation bin centers. As shown in Fig. 2, gradient magnitude $\boldsymbol{g}$ is added to the nearest n-th and (n+1)-th bin centers as $\boldsymbol{g}_n$ and $\boldsymbol{g}_{n+1}$, respectively. b_t^i of the histogram vector $\boldsymbol{v}_t$ of the t-th cell can be obtained by accumulating all the gradient magnitudes in the i-th orientation center of the t-th cell.

$$\boldsymbol{g}_n = \frac{\sin(n+1)\theta_0}{\sin\theta_0}\boldsymbol{g}_x - \frac{\cos(n+1)\theta_0}{\sin\theta_0}\boldsymbol{g}_y \tag{7}$$

$$\boldsymbol{g}_{n+1} = -\frac{\sin n\theta_0}{\sin\theta_0}\boldsymbol{g}_x + \frac{\cos n\theta_0}{\sin\theta_0}\boldsymbol{g}_y \tag{8}$$

$$b_t^i = \sum_{p_j \in C_t} g_{i,j} \tag{9}$$

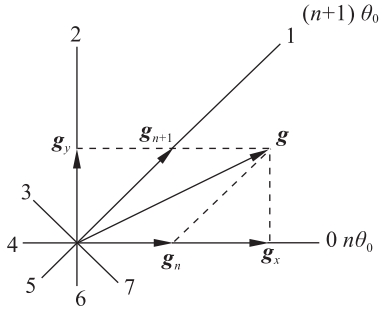


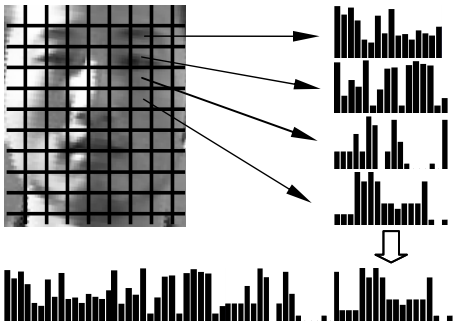
Fig. 2 Projection of gradient magnitude to the nearest orientation bin center by the parallelogram law[15]

2 Overlapped HOG

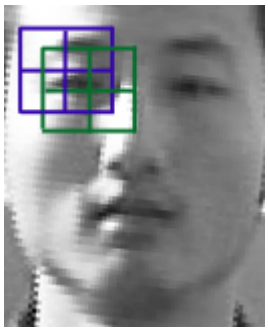
The accuracy of the eye location is very important for matching two facial images, however, 100% accuracy is impossible and the accuracy decreases dramatically when there are bad light conditions or motion blurring. The histogram itself provides some balance to this problem, but it is not enough. Thus the overlapped HOG feature was introduced to further overcome this problem. This was inspired by Dalal and Trigg's[11] conclusion that 'redundant' information introduced by overlapping significantly improves the performance for pedestrian detection, although no reason was given.

Before the overlapped HOG feature is introduced, the method for using the HOG feature must first be explained. The HOG feature is not generated for each key point here, but the entire facial image is uniformly divided into cells of the same size. The final feature of the facial image is obtained by first generating histograms of each cell and then simply concatenating them altogether (Fig. 3a). The whole process is similar to what is usually done in LBP feature extraction.

To generate an overlapped HOG feature, several of the original HOG features are first generated independently with each HOG feature produced based on a unique HOG grid. These different grids may contain



(a)



(b)

Fig. 3 (a) Image window is uniformly divided into cells of the same size with local 1-D histograms of the gradient directions accumulated and concatenated to form the final histogram feature and (b) Parts of two different grids that overlapping each other.

cells of different sizes (though in our experiments they were the same for simplicity), but they have to be placed in different locations. Thus, the cells in different HOG grids may overlap each other as in Fig. 3b. Then, either the features generated on each grid are concatenated altogether for a feature level fusion or similarity scores are calculated for the two facial images for the individual features for a score level fusion.

3 Measurement of Feature Classification Ability

According to Devuver[16], the Bayesian distance, error probability, and logarithmic information measure are related by

$$P_e \leqslant 1 - B(X|Y) \leqslant \frac{1}{2} H(X|Y) \tag{10}$$

where P_e is the Bayesian error probability, $B(X|Y)$ is the Bayesian distance, and $H(X|Y)$ is the logarithmic information measure.

From information theory,

$$P_e \leqslant \frac{1}{2}[H(X) - I(X|Y)] \tag{11}$$

Equation (11) indicates that a larger mutual information $I(X|Y)$ will reduce P_e. Let E denote the identity space and F denote the feature space. Since the entropy $H(E)$ is a constant for a given classification problem, the mutual information $I(E|F)$ can be used to evaluate the potential classification power of different features. With the homoscedastic assumption and the Gaussian distribution assumption of total scatter, $I(E|F)$ can be written as

$$I_K(E,F) = \frac{1}{2}\log_2\left(\frac{|\boldsymbol{S}_t|}{|\boldsymbol{S}_w|}\right) = \frac{1}{2}\log_2\left(\frac{|\boldsymbol{S}_b + \boldsymbol{S}_w|}{|\boldsymbol{S}_w|}\right) = \frac{1}{2}\sum_{i=1}^{K}\log_2(1+\lambda_i) \tag{12}$$

where $\boldsymbol{S}_w$ is the within-class matrix, $\boldsymbol{S}_b$ is the between-class matrix, and $\boldsymbol{S}_t$ is the total scatter matrix. λ is the associated generalized eigenvalue for the generalized eigenvalue problem used to find the Fisher basis vectors and K is the number of remaining dimensions.

The $I(E|F)$ for the different features are then compared for the classification.

4 Experimental Results

Various tests were used to study different variations of the HOG feature and to compare the potential classification power of the different features as well as their actual recognition rates on the FRGC v2.0 database and the CAS-PEAL database.

The FRGC v2.0 database[17] (Fig. 4) has more than 10 000 frontal images of 222 subjects in the training set and more than 32 000 frontal images of 466 subjects in the validation set. 10 images of each subject in the training set were randomly selected for training. The validation used the images of all 466 subjects that have expression and illumination variations. The query set was the same as that used in the standard Experiment 4 in FRGC ver2.0, which consists of 8014 single uncontrolled still images. The target set consisted of 466 single controlled still images, one for each subject, which is a little different than the target set in the standard Experiment 4 in FRGC because most practical applications do not have multiple target images for each subject, while multiple query images for each subject are common.

The CAS-PEAL database[18] (Fig. 5), obtained from the Chinese Academy of Science, was also used to evaluate the face recognition system. This database has

also been used for eye detection, facial pose estimation, and facial expression recognition. The released CAS-PEAL database contains 30 864 images of 1040 subjects (595 males and 445 females of Asians) with varying pose, expression, accessory, and lighting (PEAL). After the exclusion of large pose variations (larger than 45°), 22 020 images of all 1040 subjects were used.

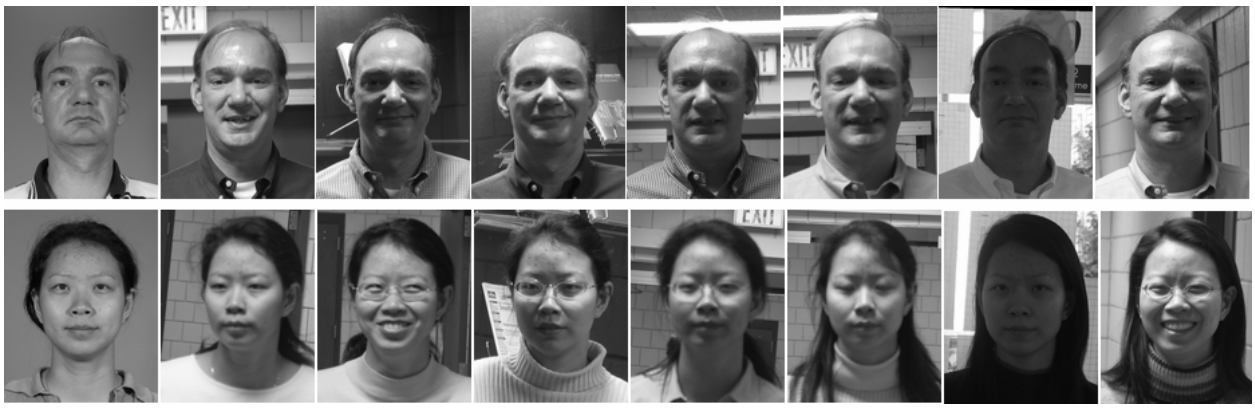


Fig. 4 Controlled and uncontrolled images in the FRGC v2.0 database

Fig. 5 Images in the CAS-PEAL database

Before the feature extraction process, all images in the FRGC or CAS-PEAL databases were cropped to smaller images having sizes of 160×130 pixels. The distances between the eyes were normalized to 60 pixels. The illuminations were normalized by simply using the arithmetic mean intensely of each image patch to normalize all the gray intensities within that patch.

The dimension of the original feature is very large (typically more than 2000). As much useful information as possible was retained to eliminate the effect of noise. PCA was used first to reduce the feature dimension. LDA was then used to extract discriminate information[5]. Finally, the nearest neighbor classifier was used for the classification due to its simplicity. In practice, if the original feature dimension is too high, the PCA eigenvalue problem might not yield a solution since the total scatter matrix may be too large. These tests had an original feature dimension limit of 6500, which means that the LBP feature could not be created with a very small cell size, since small cells would result in an LBP feature with many local histograms. Similarly, the HOG feature could not be created with a very large number of orientation bins, and the Gabor feature could not be created with many scales, even though those factors would lead to better performance. This is a tradeoff between performance and original feature length.

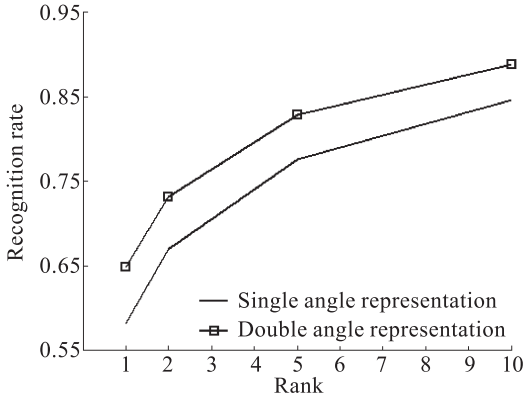
4.1 Performance of HOG feature

There are many factors that have different effects on the performance of the HOG feature for face recognition. Not all these factors affect the HOG's performance the same way as in pedestrian detection[11]. Several experiments were designed to evaluate which

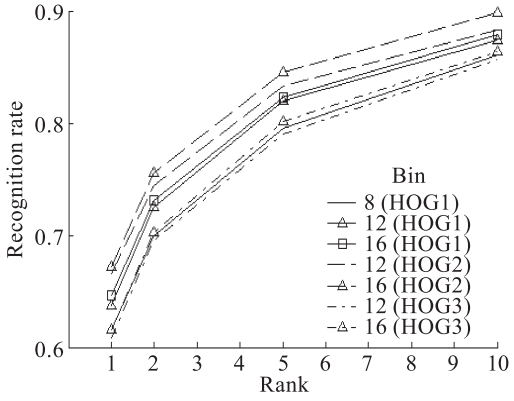
factors are critical and how they affect the recognition rate. These considered variations of the cell size, scales, orientation bins, angle representation, overlapping, etc. The cumulative recognition rates for different conditions are compared to evaluate the effects.

4.1.1 Angle representation

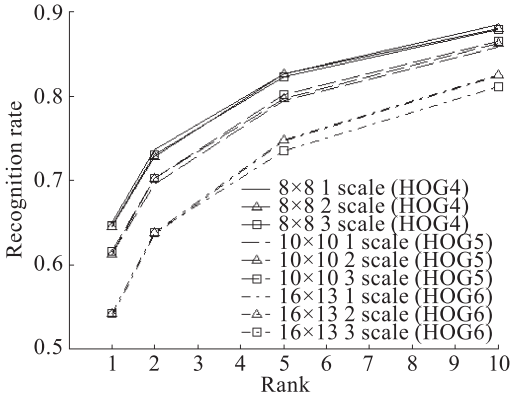
For pedestrian detection, inclusion of signed gradients (double angle representation) in the HOG descriptor decreases the performance, since "the wide range of clothing and background colors presumably makes the signs of contrasts uninformative"[11]. This is not the case for face recognition. As shown in Fig. 6a, the double angle representation leads to better performance than single angle representation. Because in face recognition, features are usually generated within a facial mask region, which eliminates the presence of background information. Inclusion of the signed gradients, more detailed information of face appearance could be obtained. Except for the way of angle representation, the HOG features in Fig. 6a were calculated with a cell size of 8×5 pixels, 12 orientation bins, one scale only, and no overlap.



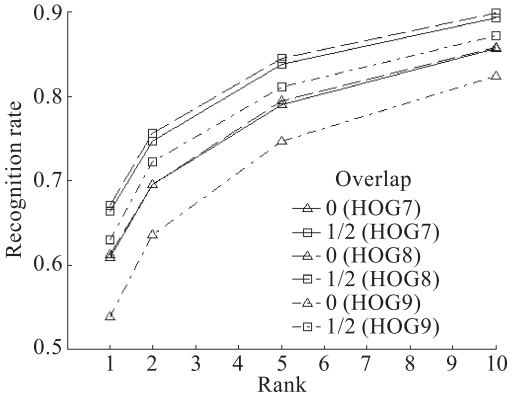
(a) Effect of the angle representation



(b) Effect of the number of orientation bins



(c) Effect of the cell size and number of scales



(d) Effect of overlap

Fig. 6 Effects of various factors on the recognition rate. (a) Including "signed" gradients (double angle representation) significantly improves the performance, (b) increasing the number of orientation bins increases performance, (c) increasing the number of scales has little effect on the performance; better performance could be achieved with relatively smaller cell size, and (d) employing overlapped cells improves the performance greatly.

4.1.2 Orientation binning

Orientation binning and votes accumulation introduce the fundamental nonlinearity to the HOG descriptor. The orientation bins are evenly spaced over 0°-360°. Each pixel calculates a weighted vote based on the orientation of the gradient element and accumulates it into the neighboring bins using the algorithm described in Section 2 over the cell the pixel belonged to. The more orientation bins, the finer structure of face pattern could be revealed by the descriptor. As shown in Fig. 6b, increasing the number of orientation bins improves performance significantly up to about 16 bins, but makes little difference beyond this. HOG1 has the following properties described below: 8×8 pixel cells, three scales, no overlapping. HOG2 is with 10×10 pixel cells, one scale only, and overlapping. HOG3 is the same with HOG1 except for its 10×10 cell size.

4.1.3 Scales and cell size

The original HOG descriptor itself did not include the concept of scaling. Inspired by the conception of multiresolution histograms[19], we tried to extract more information from face images by extending the original HOG descriptor to a multi-scale version. However, as seen in Fig. 6c, no obvious improvements are shown as the number of scales increases. Similar to orientation binning, with the smaller size of each cell, the finer structure of a face pattern could be revealed by the descriptor. As shown in Fig. 6c, significant improvements could be achieved as the cell size diminished. However, the tolerance of intra class variance, as well as the estimation of mini-pattern probability distribution also decreased as the cell size decreased, especially when there was a large number of orientation bins. In Fig. 6c, HOG4 had 16 orientation bins, 8×8 pixel cells, and no overlapping. HOG5 and HOG6 were the same with HOG4, except that they had 10×10 and 16×13 pixel cells, respectively.

4.1.4 Overlapping

These tests used overlapped cells using the method described in Section 2, with weighted impact of each pixel contributing to the histograms of at least two (could be more) cells. This brings redundant information into the HOG descriptor; however, the information is far from useless. With the overlapping, the HOG feature is more robust to small variances caused by pose or expression, at the cost of a larger feature length. Figure 6d shows that overlapping significantly increases the performance regardless of the other parameters for the HOG feature. HOG7 had 10×10 pixel cells, 12 orientation bins, and three scales. HOG8 had 10×10 pixel cells, 16 orientation bins, and only one scale. HOG9 was the same as HOG8 with 16×13 pixel cells.

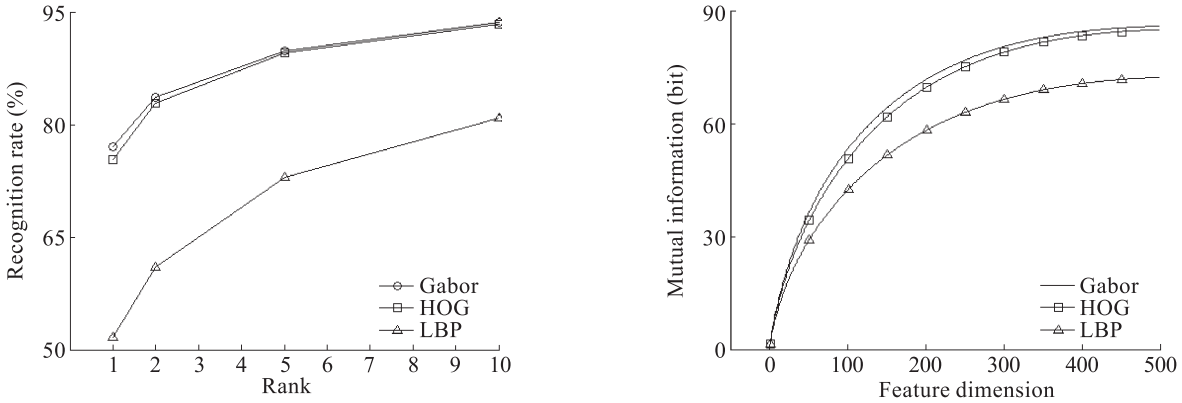
The results show that for HOG features with uniformly distributed cells on the entire image, fine scale gradients, fine orientation binning, relatively small spatial binning, and overlapped cells are all important for good face recognition performance.

4.2 Comparison of different features

This section compares the cumulative recognition rate as well as the computing time for HOG, Gabor, and LBP features.

4.2.1 Comparison of actual recognition rate

Considering the results in Section 4.1 as well as the feature length limit, the final HOG feature used for the comparisons had a 10×10 pixel cell size, 16 orientation bins, one scale, and overlapped cells. The Gabor feature used 30 filters corresponding to five different scales and six orientations. The LBP descriptor used the uniform $LBP_{(8,1)}$ (i.e., a neighborhood with eight samples and a radius of one) with a cell size of 16×13 pixels. A series of tests on the FRGC v2.0 database gave the cumulative recognition rates for the three features over a wide range of PCA and LDA dimensions (the PCA dimension ranged from 200 to 1600 with an interval of 100, the LDA dimension ranged from 80 to 220 with an interval of 20). Only the curve for the highest recognition rate of each feature is shown in Fig. 7a.



(a) Cumulative matching curves (b) Cumulative mutual information curves

Fig. 7 (a) Cumulative matching curves for the Gabor, HOG, and LBP features and (b) cumulative mutual information curves for the Gabor, HOG, and LBP features

The results in Fig. 7a show that the HOG feature achieves almost the same performance as Gabor and performs much better than the LBP feature on the FRGC v2.0 database. The computational time for the

same condition on an Intel Pentium 4, 3 GHz CPU with 1 GB RAM was 0.005 s for HOG, 0.055 s for Gabor, and 0.014 s for LBP to extract the feature for one subject on average.

4.2.2 Comparison of potential classification power

The classification powers of the different features were compared based on the mutual information for different features calculated according to Eq. (12). Since the FRGC v2.0 database was already used for the recognition rate comparison in Section 4.2.1, the CAS-PEAL database was used here to calculate the mutual information. Figure 7b shows the cumulative mutual information for different features for cut-off dimensions ranging from 0 to 500. The results show that the classification power of the Gabor feature is slightly stronger than that of HOG, and much stronger than that of LBP. This conclusion is in accord with the conclusions from the comparison of these actual recognition rates in Section 4.2.1, even though the CAS-PEAL and FRGC v.2.0 databases are very different (e.g., one having only Asian facial images while the other mainly consists of western facial images). This validates the criterion developed to evaluate the classification power of the different features.

5 Conclusions

As a relatively simple local descriptor, the HOG feature is widely used in applications like pedestrian detection and tracking, but has rarely been used in face recognition. A fast computational method was developed and many different factors that affect the HOG's performance were evaluated to develop a HOG descriptor with fine-scale gradients, fine orientation binning, relatively small spatial binning (cell size), and overlapped cells over the entire image which succeeded in achieving almost the same performance but with a lower time cost compared to the Gabor descriptor, and better accuracy than the LBP descriptor.

The use of mutual information as a general measurement of potential classification power for different features may be more theoretically important here than the successful application of the HOG feature in face recognition. The mutual information enables comparison among different kinds of features regardless of the specific application environment.

This in-depth exploration of the HOG features in face recognition will be extended in future work concentrating on the fusion of the HOG feature with other features, since extra features improve robustness with correct matches, and do little harm other than their cost of computation. Relatively simple but effective features like the HOG features are the best choices for fusion tasks.

References

[1] Turk M, Pentland A. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 1991, **3**(1): 71-86.

[2] Freeman W, Adelson E. The design and use of steerable filters. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1991, **13**(9): 891-906.

[3] Belongie S, Malik J, Puzicha J. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2002, **24**(4): 509-522.

[4] Zhao W, Chellappa R, Rosenfeld A, et al. Face recognition: A literature survey. *ACM Comput. Surv.*, 2003, **35**: 399-458.

[5] Belhumeur P N, Hespanha J P, Kriegman D J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1997, **19**(7): 711-720.

[6] Ahonen T, Hadid A, Pietikainen M. Face recognition with local binary pattern. In: Proc. 8th Eur. Conf. Computer Vision. Prague, Czech, 2004: 469-481.

[7] Gabor D. Theory of communication. *Journal of Institute for Electrical Engineering*, 1946, **93**(Ⅲ): 429-457.

[8] Lades M, Vorbruggen J C, Buhmann J, et al. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Computers*, 1993, **42**(3): 300-311.

[9] Wiskott L, Fellous J M, Kruger N, et al. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1997, **19**(7): 775-779.

[10] Freeman W T, Roth M. Orientation histograms for hand gesture recognition. In: Intl. Workshop on Automatic Face and Gesture Recognition. IEEE Computer Society, Zurich, Switzerland, 1995: 296-301.

[11] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Diego, CA, USA, 2005, **1**: 886-893.

[12] Lowe D G. Object recognition from local scale-invariant features. In: Proceedings of the 7th International Conference on Computer Vision. Kerkyra, Greece, 1999, **2**: 1150-1157.

[13] Granlund G H. In search of a general picture processing

operator. *Computer Graphics and Image Processing*, 1978, **8**(2): 155-173.

[14] Liu C L, Nakashima K, Sako H, et al. Handwritten digit recognition: Investigation of normalization and feature extraction techniques. *Pattern Recognition*, 2004, **37**(2): 265-279.

[15] Liu Hailong. Offline handwritten character recognition based on descriptive model and discriminative learning [Dissertation]. Tsinghua University, Beijing, China, 2006.

[16] Devuver P A. On a new class of bounds on Bayes risk in multihypothesis pattern recognition. *IEEE Trans. Computers*, 1974, **C-23**(1): 70-80.

[17] Phillips P J, Flynn P J, Scruggs T, et al. Overview of the face recognition grand challenge. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Diego, CA, USA, 2005, **1**: 947-954.

[18] Gao Wen, Cao Bao, Shan Shiguang, et al. The CAS-PEAL large-scale Chinese face database and baseline evaluations. Technical Report JDL-TR-04-FR-001. Joint Research and Development Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, 2004.

[19] Hadjidemetriou E, Grossberg M D, Nayar S K. Multiresolution histograms and their use for recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2004, **26**(7): 831-847.