Neurocomputing 194 (2016) 10–23

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Hybrid human detection and recognition in surveillance

Qiang LIU [a,b], Wei ZHANG [a,*], Hongliang LI [c], King Ngi NGAN [b]

[a] School of Control Science and Engineering, Shandong University, China
[b] Department of Electronic Engineering, The Chinese University of Hong Kong
[c] School of Electronic Engineering, University of Electronic Science and Technology of China

ARTICLE INFO

Article history:
Received 13 May 2015
Received in revised form
2 February 2016
Accepted 12 February 2016
Communicated by Shiguang Shan
Available online 19 February 2016

Keywords:
Head–Shoulder Detector
Human recognition
AdaBoost
Overlapping Local Phase Feature
Gaussian Mixture Model
Surveillance

ABSTRACT

In this paper, we present a hybrid human recognition system for surveillance. A Cascade Head–Shoulder Detector (CHSD) with human body model is proposed to find the face region in a surveillance video frame image. The CHSD is a chain of rejecters which combines the advantages of Haar-like feature and HoG feature to make the detector more efficient and effective. For human recognition, we introduce an Overlapping Local Phase Feature (OLPF) to describe the face region, which can improve the robustness to pose change and blurring. To well model the variations of faces, an Adaptive Gaussian Mixture Model (AGMM) is presented to describe the distributions of the face images. Since AGMM does not need the facial topology, the proposed method is resistant to face detection error caused by imperfect localization or misalignment. Experimental results demonstrate the effectiveness of the proposed method in public dataset as well as real surveillance video.

1. Introduction

Nowadays, surveillance cameras are deployed almost every corner and street over the world, especially in big cities, to watch and manage the activities of human being. For example, there are around 500,000 CCTV cameras in London and 4,000,000 cameras in UK [1]. It is impossible to hire enough security guard to monitor the huge number of cameras constantly, 24 h and 7 days. Generally, the camera feeds are recorded without monitoring and the videos are mainly used for a forensic or reactive response to crime or terrorism after the event happened. However, only recording surveillance video is not enough to prevent the terrorists. Intelligent detection of events and persons of interest from the camera feeds before any attack happens is urgently required for surveillance purpose.

As an intelligent surveillance system, it should be able to identify where and who is in the scene. An intelligent surveillance system mainly includes human detection and recognition. However, in practice, it is very challenging to find and recognize human when illuminations, expressions, and poses vary. Besides, surveillance videos also have low quality due to the long distance of the target from the camera, out-of-focus blur or motion blur caused by motion between the target and camera, or a combination of all factors aforementioned. Besides, camera noise and image distortion incurred by optical sensor and network transmission also affect the performance of human detection and recognition.

In the surveillance human recognition literature, most work was presented with the assumption that the face detection is given. To deal with pose variation, Gaussian mixture Models [2,3] are learned from training data to characterize human faces, head pose variations, and surrounding changes. In [4,5] use 3D model to aid face recognition to robust to facial expression and pose variations and further improvement by adding auxiliary information, such as motion and temporal information between frame images. And [6] uses “Frontalization” face to do face recognition and gender estimation. Ma et al. [7] improved the accuracy of pose estimation by investigating the symmetry property of the face image. To deal with the illumination variations, Thermal Infrared Sensor (TIRS) [8] was used to measure energy radiations from the object, which is less sensitive to illumination changes. However, thermal images have low resolution and are unable to provide rich information of the facial features. To account for blurring problem, Hennings-Yeomans et al. [9] first performed restoration to obtain images with better quality [10], and then fed them into a recognition system. Rather treating restoration and recognition separately, Zhang et al. [11] proposed a joint blind restoration and recognition model based on sparse representation to deal with frontal and well-aligned faces. Grgic et al. [12] also provided a surveillance face database collected in uncontrolled indoor environment using five types surveillance cameras of various qualities and applied principal component analysis (PCA) for face recognition. In [13,14], each face was described in terms of multi-region modelled by probabilistic distributions, such as GMMs, followed by a normalized

* Corresponding author.
E-mail address: davidzhangsdu@gmail.com (W. ZHANG).

http://dx.doi.org/10.1016/j.neucom.2016.02.011
0925-2312/

distance calculated between two faces, which can be efficient to deal with faces with illumination and misalignment. However, face recognition is still an open problem in surveillance, although techniques [15–17] used in face recognition literature [18–20,74] perform well with the cooperative subjects in controlled applications. Also, current face detectors are unable to find the face well in the low-quality surveillance video.

In this paper, we present a hybrid human recognition system by integrating face detection and recognition together as shown in Fig. 1. For face detection, we propose to find the Head–Shoulder (HS) region first by the Cascade Head and Shoulder Detector (CHSD), and then employ the trained body model to get the face region for recognition. In face recognition, to represent face region discriminatively, we propose an Overlapping Local Phase Feature (OLPF) which is robust to image blur and pose variation without adversely affecting discrimination performance. To model faces robustly, a Fixed Adaptive Gaussian Mixture Model (FGMM) is developed to describe the distribution of the face data, but FGMM may be degraded because of different subjects needing different numbers of Gaussians to model the variations of faces. Therefore, an Adaptive Gaussian Mixture Model (AGMM) is proposed to optimally build the model for each subject. Without face topology, the proposed AGMM is insensitive to the initial face detection without alignment. Combining AGMM and OLPF, our method can handle faces with multiple uncontrolled issues in surveillance, such as misalignment, pose variation, illumination changing, and blurring. The proposed detection and recognition scheme can be extended to other objects of interest with similar properties such as cars and animals.

The organization of the paper is arranged as follows: In Section 2, we give the structure of CHSD and the details of how to train each filter in the CHSD. The proposed face recognition algorithm are discussed in Section 2.1. Extensive experiments are given in Section 2.2 to demonstrate the robustness of our method. Conclusions are summarized in Section 2.2.1.

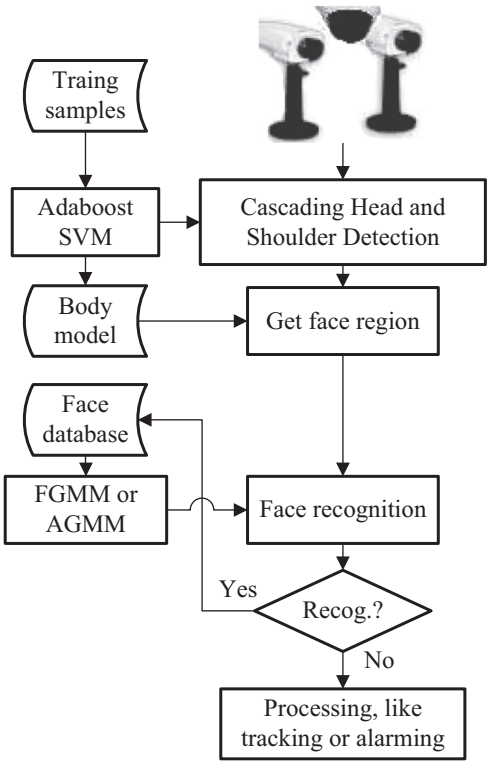


Fig. 1. Diagram of proposed system.

2. Cascade Head and Shoulder Detection

As aforementioned, in general surveillance condition, people and the target scene cannot be strictly controlled. The face to be recognized may not appear as assumed in [11,19], such as the frontal face with proper lighting. So the captured faces may differ substantially in pose, illumination and expression. Some examples are given in Fig. 2 from an indoor surveillance application to show the variations of pose and illumination in the face region. For these cases, traditional face detector [17,21] may not work well to locate the face region effectively and correctly. To overcome these problems in unconstrained conditions, we propose to detect HS region first, and then use the human body model to obtain the face region.

The proposed method is inspired by [22,23] with the use of a dense grid of Histograms of Oriented Gradients (HoG) and linear Support Vector Machine (SVM) to detect human. However, we found that those detectors are not enabled to allow fast rejection in the early stages. It works slowly and can only process 320×240 images at 10 frame per second (fps) in a sparse scanning manner. In this paper, we intend to speed it up to real-time without quality loss by cascading new classifiers.

The idea of CHSD is to use a cascade of rejecters to filter out a large number of non-HS samples while preserving almost 100% of HS regions. Thus the number of candidates can be reduced significantly before more complex classifiers are called upon to achieve low false positive rates. As shown in Fig. 3(a), CHSD includes three parts: initial feature rejecter, Haar-like rejecter, and HoG classifier.

2.1. Initial feature rejecter

In this rejecter, one of the features is the regional variances which can be obtained by limited computations[1] from two integral images, i.e., integral image and integral image of the squared image. Those integral images will also be used to perform illumination normalization in the preprocessing step and feature calculation in the Haar-like rejecter, so no additional computation is required in this rejecter. Assuming that σ_k denotes the variance of the kth region, our training process is described in Algorithm 1.

The other feature of the first rejecter is the difference between two blocks no matter whether they are adjacent or not. The training method in Algorithm 2 is similar to that in Algorithm 1 with a few minor modifications from steps (a)–(c).

Algorithm 1. Training for rejecter using variance features.

1. Input training data $(\langle x_1, y_1\rangle, \ldots, \langle x_n, n\rangle)$ where $y_i \in \{0, 1\}$ for non-HS and HS regions,respectively.
2. Initialize rejecter label $l_i = 0$, for $y_i = 0$;
3. For $t = 1, \ldots, T$:
   a. Find the minimal and maximal values of σ_k for each region k from the training samples, which are denoted by ${\sigma_k}^{\min}$ and ${\sigma_k}^{\max}$, respectively.
   b. Compute the rejection number r_k for non-HS training samples, with a parity p adjusting the in-equality direction: $r_k^p = \sum_{y_i = 0, l_i = 0} sign\,|\,p\sigma_{i,k} > p\sigma_k^p\,|$, $p = -1$ for ${\sigma_k}^{\min}$ and $p = 1$ for ${\sigma_k}^{\max}$
   c. Choose the region with the highest rejection number
   d. Set label $l_i = 1$ for all rejected sample $\{i\}$.
4. Output the combined classifiers.

[1] Any two-rectangle feature can be computed in six array references, any three-rectangle feature in eight, and any four-rectangle feature in just nine.

Fig. 2. Examples of images from surveillance videos (4CIF).

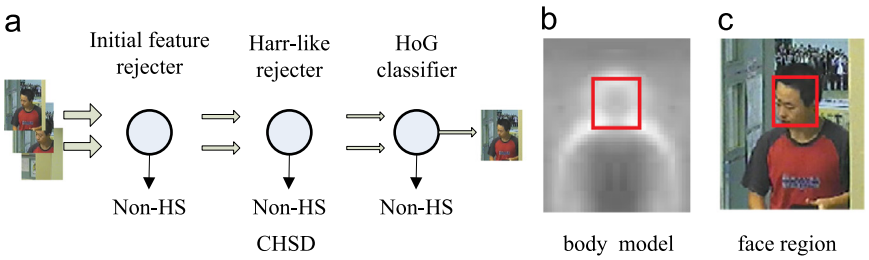


Fig. 3. Face detection use CHSD and body model. (a) The structure of CHSD. (b) Body model and trained face region marked by the red rectangle. (c) Detected face region. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

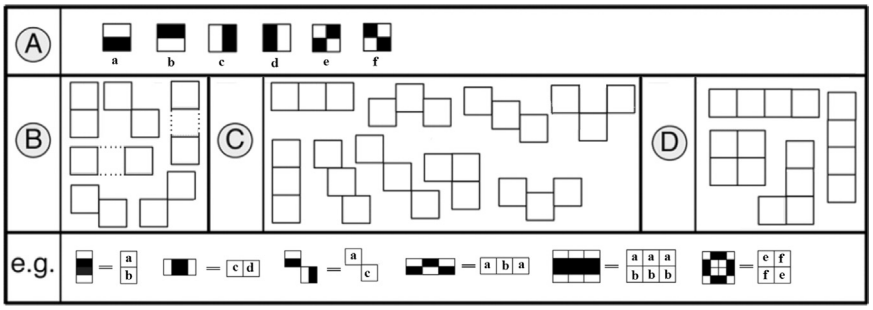


Fig. 4. Examples of the proposed Haar-like features.

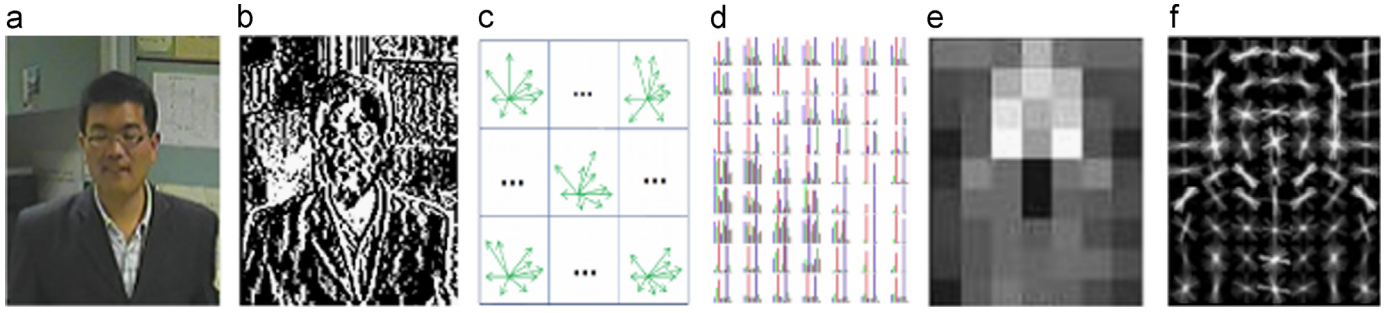


Fig. 5. HOG extraction and the SVM training results. (a) A test image. (b) Gradient image of the test image. (c) Orientation and magnitude of Gradient in each cell. (d) HoG of cells. (e) The weights of positive SVM in the block. (f) The HoG descriptor weighted by the positive SVM weights.

In the initial feature rejecter, the characteristics of the variance and block difference of the image segments are used to form a rejecter. It demonstrates that even simple features can be used to construct an efficient rejecter. Since these features are also used by Haar-like features in the following rejecter, in some sense, no additional computation is needed for feature generation in the initial feature rejecter.

2.2. Haar-like rejecter

For a candidate window accepted by the initial feature rejecter, it will be further evaluated by the learning based Haar-like rejecter. In this part, we present how to construct a strong rejecter using the Haar-like features trained by AdaBoost method.

2.2.1. Feature

The simple Haar-like features, shown as Fig. 4(A), have been successfully applied to face detection by Viola and Jones based on a fast calculation method [15]. Some simpler features, i.e., the colour relationship between two pixels, were used to perform sex identification [25]. In order to improve the performance, more rotated Haar-like features and scalar Haar-like features were extended in [26] and [27] to deal with in-plane rotations and multi-view face detection, respectively.

Most previous methods construct the weak classifier using boosting features from the huge number of feature sets represented by Haar-like features. In [28], a template pool is generated by sliding bounding boxes of different sizes over the pre-defined pedestrian body shape model. In our feature pool, the model is designed based on

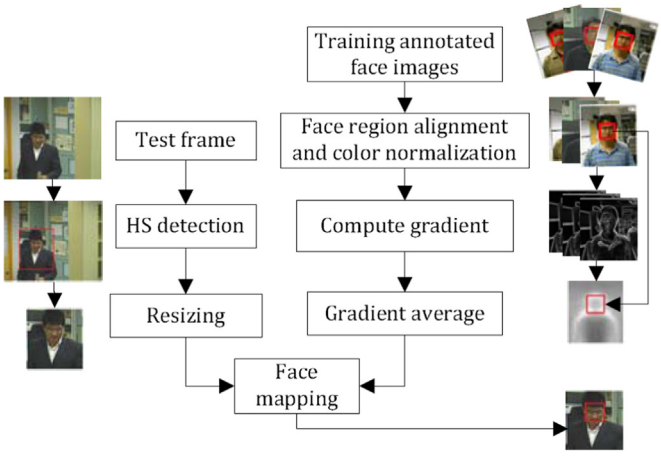


Fig. 6. Diagram of the body model generation.

the properties of Head and Shoulder, i.e., the shape information and pixel intensity in LUV. The training is performed via AdaBoost to boost up the most informative feature for classification. To improve the performance of the weak classifier, joint Haar-like features [29] and filtered low-level features [30] are employed. The examples in the last row of Fig. 4 are the features generated by combining the basic Harr-like features shown in Fig. 4(A) based on the patterns shown in Fig. 4 (B)–(D). In some sense, the joint Haar-like features are more like the "toy bricks" which can be built according to a certain composition. Each feature consists of multiple "bricks" which are combined by means of addition, subtraction, and absolute value operations.

For a window with the size of 64×80 and the scale$=1$, the total number of features is quite large, e.g., 32,879 for feature A, 554,034 for feature B, 713,412 for feature C, and 106,641 for feature D. The features with the best classification of the training dataset will be boosted from the tens of millions of the features to construct a rejecter.

2.2.2. Training

There are many boosting approaches [31] for object classification by machine learning, such as AdaBoost [15,26,32], FloatBoost [27], Kullback-Leibler Boosting [33]. In our previous work [34], we used AdaBoost algorithm for training face detector. It is known that AdaBoost approach can be interpreted as a greedy feature selection process by which a small set of features and associated cascade weights are selected with the lowest classification errors.

Algorithm 2. Training for rejector using block difference.

a. Find the minimal and maximal values of $D_{k,j} = M_k - M_j$ for two arbitrary blocks from the training samples, which are denoted by $D^{min}_{(k,j)}$ and $D^{max}_{(k,j)}$, respectively. M is the mean value of the given block.
b. Compute the rejection number $r_{k,j}$ for non-HS training samples, with a parity p adjusting the inequality direction: $r^p_{(k,j)} = \sum_{y_i=0, l_i=0} sign|pD_{i,(k,j)} > pD^p_{(k,j)}|$, $p=-1$ for $D^{min}_{(k,j)}$ and $p=1$ for $D^{max}_{(k,j)}$
c. Choose the region with the highest rejection number.

Haar-like rejecter is considered strong because it is a weighted combination of many weak rejecters. Although each weak rejecter constructed by one feature cannot provide good rejection for the training samples, the appropriate combination of them with weighting can improve the performance of the final classification significantly, which is described in Algorithm 3.

2.3. HoG feature classifier

Viola et al. [35] built an efficient moving pedestrian detector in a surveillance environment using AdaBoost to train a cascade rejecter based on the Haar-like features and spatial differences. But the detection performance relies significantly on the available motion information. Dalal and Triggs [22] proposed a human detection algorithm with a dense grid of Histograms of Oriented Gradients (HoG) features which have been proved to be more powerful than the Haar-like features in human detection. In [30], Zhang et al. used HOG+LUV as low-level features, while adding optical flow features to do human detection. In our system, we focus on detecting the HS region with the assumption that the HS region is fully visible. In proposed CHSD, the HoG feature is employed in the final classifier as benchmark.

Algorithm 3. AdaBoosting training

1. Input training data $(\langle x_1, y_1\rangle, \ldots, \langle x_n, n\rangle)$ where $y_i \in \{0,1\}$ for non-HS and HS regions, respectively.
2. Initialize sample weights $\omega_{1,i} = (1/2p), (1/2q)$ where p and q are the number of positive and negative samples.
3. For $t = 1, \ldots, T$:
   a. Normalized weights $\omega_{t,i}$.
   b. Compute the classification error for each feature f using $\epsilon_f = \sum_i \omega_{(t,i)} |h(f, x_i) - y_i|$
   c. Choose the best weak classifier $h_t(x)$ with the lowest error ϵ_t .
   d. Update weight. $\omega_{(t+1,i)} = \omega_{(t,i)} \left(\frac{\epsilon_t}{1-\epsilon_t}\right)^{1-|h_i(x_i)-y_i|}$
4. Output the combined classifiers.

$$h(x) = \begin{cases} 1, \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0, \text{otherwise} \end{cases}$$

2.3.1. Features

To extract HoG feature from an image, such as Fig. 5(a), it is divided into uniformly sized cells and a group of cells is integrated into a block in a sliding fashion with blocks overlapping with each other vertically and horizontally shown as in Fig. 5(c). Each cell from its gradient image is quantized and projected to a 9-bin Histograms of Oriented as in Fig. 5(d). The feature representing a detection window is a concatenated vector of all its cells and then normalizes to a $\mathcal{L}_2$-*norm* unit length. These feature vectors are then classified by a linear Support Vector Machine (SVM).

2.3.2. Training

The training data consists of a large set of images with bounding boxes around each instance of an object. We reduce the problem of learning to a binary classification problem. Let $(\langle x_1, y_1\rangle, \ldots, \langle x_n, n\rangle)$ be a set of labelled examples where $y_i \in \{-1, 1\}$ and x_i specifies a HoG feature of a training image. We construct a positive example from each bounding box in the training set. Negative examples come from images that do not contain the target object. A soft $(C = 0.01)$ linear SVM is trained with the SVMLight [36] algorithm. The objective function is then increased by a function which penalizes non-zero ξ_i for each sample, and the optimization becomes a trade off between a large margin, and a small error penalty. If the penalty function is linear, the optimization problem becomes:

$$\min_{\omega,\xi,b} \left\{ \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{n} \xi_i \right\} \quad (1)$$

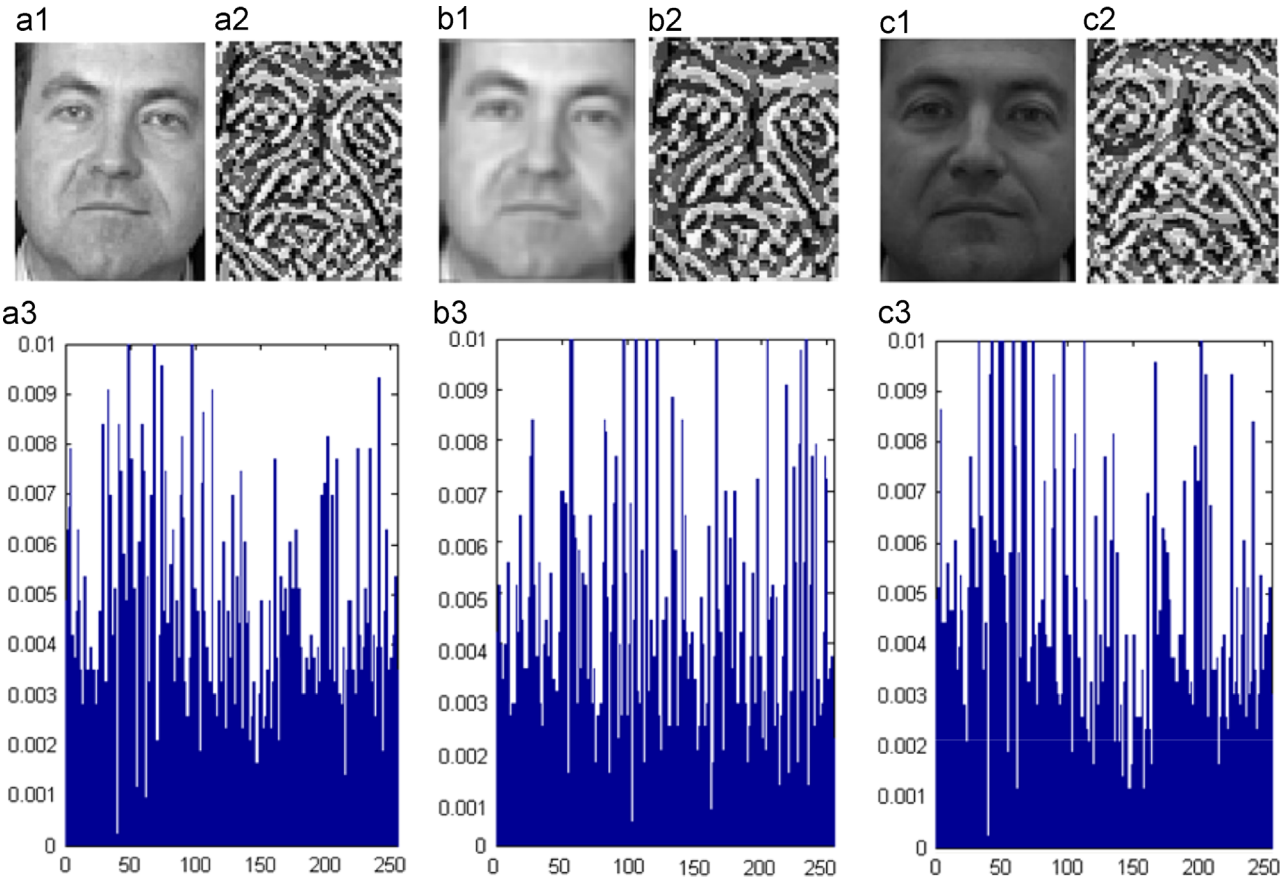


Fig. 7. Invariant property of phase feature to blur and illumination. a, b, and c show the original image, images with blurring ($\sigma = 2$), and image with different illumination. 1, 2, and 3 denote the original image, the LPF image, and the LPF histogram. The Bhattacharyya distances between two histograms are 0.0846 (a3 and b3) and 0.1035 (a3 and c3).

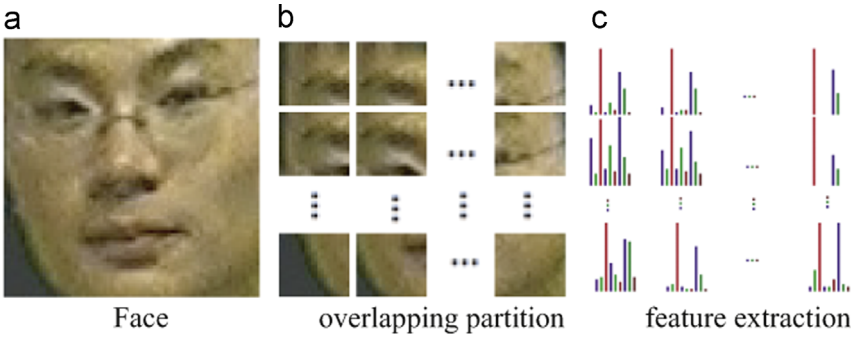


Fig. 8. OLPF feature extraction.

subject to $y_i(\omega^T \cdot x_i - b) \geq 1 - \xi, \xi \geq 0$ for any $i = 1, \ldots, n$. The training results and weighted HoG are shown in Fig. 5(e) and (f).

2.4. Cascade of classifiers

In CHSD, multiple layers of cascade classifiers is employed to reject as many non-HS samples as possible at the earliest stages with limited computation, which reduces the detection time greatly. The first layer is the initial feature rejecter where common features like variance and difference are used and calculated efficiently from the integral images. The second layer is the Haar-like rejecter constructed by a cascade of Haar-like features. In this rejecter, every weak rejecter is adjusted to have a very high detection rate (e.g., 99.9%), but a moderate false positive rate (50%) after the AdaBoost learning. If 10 of the above rejecters are bounded together, the false alarm rate and detection rate would be 9.7×10^{-4} and 0.99, respectively. The first two rejecters get rid of the majority of the non-HS samples while retaining the detection rate of almost 100%. The last layer is the HoG classifier which only needs to deal with tens of HS candidates for an image. So the classification can be finished quickly even for high dimensional data (2268 dimensions).

To generate a body model, we randomly select 2000 samples and annotate the face regions for training. As illustrated in the rightmost column of Fig. 6, HS regions are aligned with annotated face region and cropped to the same size 64×80. More details can be found in the Preprocessing Section 4.1. After colour normalization, the HS gradients are calculated using Sobel filter. The body model is produced by combining those gradient images and face region is annotated according to the annotation of aligned training samples.

As shown in the leftmost column, in face detection, the input frame is first fed into CHSD to gain the Head–Shoulder region. Then, the trained body model is mapped on it to finally extract the face region.

3. Face recognition

In traditional face recognition algorithms, from utilizing the facial properties and relationships, such as areas, distances, and angles to projecting the face image to feature spaces, e.g., Eigenface [37], Fisherface [38], Laplacianface [39] and derivative domain [40,41], those methods were designed for well aligned, uniformly illuminated, and frontal pose face images. While, in practice, it is almost impossible to satisfy these requirements, especially in security surveillance system. Consequently, many efforts have been made to develop algorithms for unconstrained face images [42,43]. Instead of using global features, local appearance descriptors such as Gabor jets [44], Local Binary Patterns [45], SIFT [46], HOG [47] and SURF [48] were employed because of their robustness to occlusion, expression, pose and smaller sample size than the global feature.

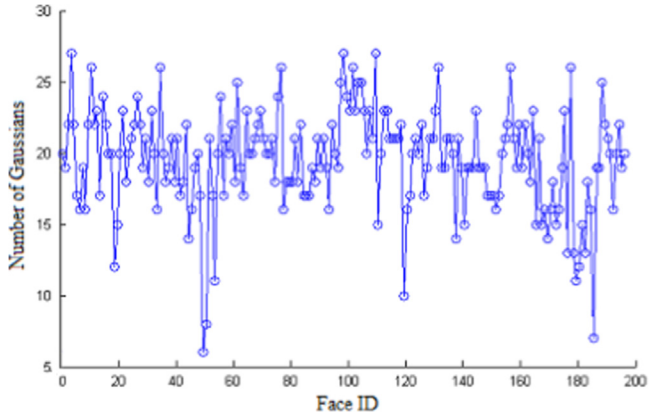


Fig. 9. Number of GMMs for each subject in FERET.

To mitigate against the low-resolution and blurring problems that often suffered in the surveillance images, Hennings-Yeomans et al. [49] proposed a method to extract features from both the low-resolution faces and their super-resolution ones within a single energy minimization framework. On the other hand, Gupta et al. [50] alternated between recognition and restoration with the assumption of a known blurring kernel. And Nishiyama et al. [51] proposed to improve the recognition of blurry faces with a predefined finite set of blurring kernels. Using the theory of sparse representation and compressed sensing, Wright et al. [52] yield new insights into two crucial issues in face recognition: the role of feature extraction and the difficulty of occlusion.

For the above methods, alignment is an indispensable preprocessing step, i.e., fix the coordinates of corners (e.g., eyes, nose) and then normalize to the same scale. However, it is known that automatic alignment is still a challenging problem for real-time system. Especially, faces detected automatically are often unsatisfactory at different scales and locations. Even detecting faces in surveillance image is a challenging task because of the highly uncontrolled pose, non-uniform illumination, camera noise, and compression distortion in network transmission. However, those constraints are relaxed in the proposed face recognition algorithm because of distinctive feature representation and robust face model for face region, which will be investigated in the following sections.

3.1. Overlapping Local Phase Feature (OLPF)

Local Binary Pattern (LBP) as a local feature has been proven to be highly discriminative descriptors for various applications, including image retrieval, surface inspection, texture classification and segmentation. However, most LBP-based algorithms [13,45] use a rigid

Fig. 10. Evaluation of the robustness of AGMM (the normalized similarity probability is given below the image).

descriptor matching strategy that is sensitive to pose variation and misalignment of face, and thus cannot work well in surveillance. In this section, we propose a modified LBP-like feature, Overlapping Local Phase Feature (OLPF), to overcome the difficulties of unconstraint face recognition in surveillance.

In the traditional methods, they may be robust to illumination or expression but may not be efficient to a blurred image, like PCA [37] and LBP [45]. We propose the OLPF based on the phase feature [53] which is extracted from the frequency domain by Fourier Transform. In mathematical formulation, the image blurring process in the time domain can be described as:

$$b(\mathbf{x}) = (i * k)(\mathbf{x}) \tag{2}$$

where $i(\mathbf{m})$ is the original image, $b(\mathbf{m})$ is the observed blurred image, and $k(\mathbf{m})$ is the blurring kernel, in the time domain. $*$ denotes 2D convolution and $\mathbf{m}$ is a vector of coordinates $[m, n]^T$. In the Fourier domain, (2) corresponds to

$$\mathcal{B}(\mathbf{u}) = (\mathcal{I} \cdot \mathcal{K})(\mathbf{u}) \tag{3}$$

where $\mathcal{B}(\mathbf{u})$, $\mathcal{I}(\mathbf{u})$ and $\mathcal{K}(\mathbf{u})$ are the discrete Fourier transforms (DFT) of the blurred image $b(\mathbf{m})$, the original image $i(\mathbf{m})$, and the blurring kernel $k(\mathbf{m})$, respectively, and $\mathbf{u}$ is a vector of coordinates $[u, v]^T$. We may separate the magnitude and phase parts of (3) into

$$|\mathcal{B}(\mathbf{u})| = |(\mathcal{I}(\mathbf{u})| \cdot |\mathcal{K}(\mathbf{u})| \text{ and } \angle\mathcal{B}(\mathbf{u}) = \angle\mathcal{I}(\mathbf{u}) + \angle\mathcal{K}(\mathbf{u}) \tag{4}$$

If the blurring kernel $k(\mathbf{m})$ is assumed to be centrally symmetric, namely $k(\mathbf{m}) = k(-\mathbf{m})$, its Fourier transform is always real-valued $\mathcal{K}(\mathbf{u}) = \mathcal{R}e\{\mathcal{K}(\mathbf{u})\}$, and as a consequence its phase part is only a two-valued function, given by

$$\angle\mathcal{K}(\mathbf{u}) = \begin{cases} 0, & \text{if} \mathcal{K}(\mathbf{u}) \geq 0 \\ \pi, & \text{if} \mathcal{K}(\mathbf{u}) < 0 \end{cases} \tag{5}$$

This means that $\angle\mathcal{B}(\mathbf{u}) = \angle\mathcal{I}(\mathbf{u})$ for all $\mathcal{K}(\mathbf{u}) \geq 0$, therefore a blur invariant representation can be obtained from the phase part.

The frequency could be computed using a short-term Fourier transform (*STFT*) on $M \times M$ neighbourhoods $\mathcal{N}_{\mathbf{m}}$ at each pixel position $\mathbf{m}$ of the image $i(\mathbf{m})$ defined by

$$\mathcal{I}^{\mathcal{N}}(\mathbf{u}, \mathbf{m}) = \sum_{\mathbf{y} \in \mathcal{N}_{\mathbf{m}}} i(\mathbf{y})\, r(\mathbf{y} - \mathbf{m}) e^{-j2\pi \mathbf{u}^T \mathbf{y}} \tag{6}$$

where $r(\mathbf{m})$ is a rectangle window function defining the neighbourhood $\mathcal{N}_{\mathbf{m}}$ of $\mathbf{m}$. The transform can be efficiently evaluated for all image position $\mathbf{m} \in \{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_{\mathcal{N}}\}$ using 1-D convolutions for the rows and columns successively. The local Fourier coefficients are computed at four frequency point $\mathbf{u}_1 = [a, 0]^T$, $\mathbf{u}_2 = [0, a]^T$, $\mathbf{u}_3 = [a, a]^T$, and $\mathbf{u}_4 = [a, -a]^T$, where a is a sufficiently small scalar to satisfy $\mathcal{K}(\mathbf{u}_i) > 0$. As a invariant feature to blur $\mathcal{I}^{\mathcal{N}}_{\mathbf{m}}$ is extracted by observing the signs of the real and imaginary parts of each component in the Fourier domain for recognition. A LBP-like method quantizes the phase information:

$$q_j = \begin{cases} 1, & \text{if } g_j(\mathbf{m}) \geq \epsilon \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $g_j(\mathbf{m})$ is the jth component of the vector $\mathcal{G}_{\mathbf{m}} = [\mathcal{R}e\{\mathcal{I}^{\mathcal{N}}_{\mathbf{m}}\}, \mathcal{I}m\{\mathcal{I}^{\mathcal{N}}_{\mathbf{m}}\}]$. ϵ is a robust threshold which we introduce to control the quantization degree. The resulting eight binary coefficients $q_j(\mathbf{m})$ (8-neighbourhood) are represented as integer values between 0 and 255 using binary coding

$$f_{LPF} = \sum_{j=1}^{8} q_j(\mathbf{m}) \cdot 2^{j-1} \tag{8}$$

An example in Fig. 7, the original image (a1), the blurred image (b1), and different illumination image (c1), is represented by the quantized phase histograms as shown in (a3), (b3), and (c3). And their phase image are described in Fig. 7(a2), (b2), and (c2). From the Bhattacharyya distance measuring the similarity of between two quantized histograms, it is obvious that the extracted phase feature can tolerate with severe blurring and illumination changing.

Head pose is believed to be one of the hardest problems for face recognition [54]. Although phase feature can tolerate with blurred image and poor illumination, it is sensitive to the pose variation and misalignment usually happened in surveillance. Inspired by the "bag-of-feature" approach [55], we develop an Overlapping Local Phase Feature (OLPF), which describes a face as a set of temporally correlated feature vectors as shown in Fig. 8. For each face, we first divide it into small, uniformly sized, overlapped blocks as shown in Fig. 8(b). Then descriptive features (Fig. 8(c)) are extracted from each block to form a vector which is used to perform training and recognition. The robustness to pose variations is attributed to the explicit allowance for movement of face areas, when comparing face images of a particular person at various poses. Changes occurring at one facial component (e.g., the mouth) only affect the subset of face areas that cover this particular component. Therefore, OLPF-based face descriptor is not only robust to blurring but also to pose, expression, and misalignment.

3.2. Fixed Gaussian Mixture Model (FGMM)

In surveillance system, it is difficult to get an ideal frontal face image, because the cameras are normally mounted under the ceiling where subjects rarely pose for. Although face synthesis algorithm like that described in [6] can convert the lateral faces to frontal ones, the synthesized faces still have residual artefacts which may degrade the recognition performance significantly. In [55] , a "bag of features" approach was shown to perform well in the presence of pose variations. It is based on dividing the face into overlapping uniform-sized blocks, analysing each block with the Discrete Cosine Transform (DCT) and modelling the resultant set of features via a Gaussian Mixture Model (GMM). In our face recognition, OLPF is employed to replace the DCT feature. Given a face image, it is normalized to the size of 64×80 pixels and a 1073×64 feature matrix is obtained to represent the face with the blocksize of 8×8 and 4 overlapping pixels. By assuming that the feature vectors $\mathcal{X}$ are independent and identically distributed (i.i.d.), the likelihood of it belonging to the person i is

$$P(\mathcal{X} | \lambda^{[i]}) = \prod_{n=1}^{N} P(x_n | \lambda^{[i]}) = \prod_{n=1}^{N} \sum_{g=1}^{G} \omega_g^{[i]} \mathcal{N}(x_n | \mu_g^{[i]}, \Sigma_g^{[i]}) \tag{9}$$

Fig. 11. Colour normalization (a are the original images; b are the normalized ones).

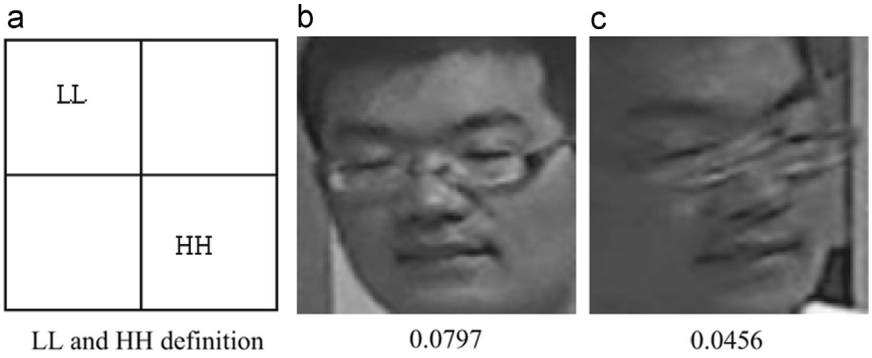


Fig. 12. Blurring image removed by HFT (0.5 was used as the setting in the experiments).

Fig. 13. Examples of the positive and negative training data.

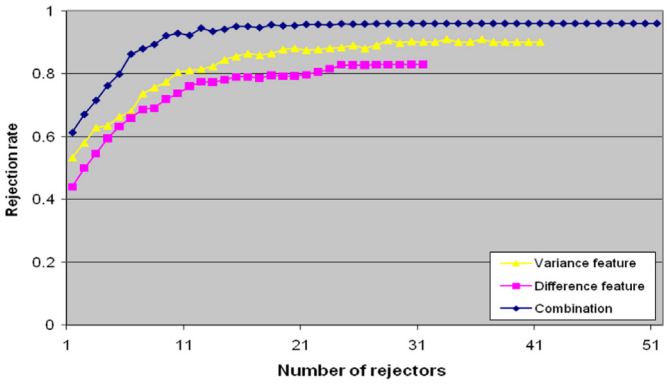


Fig. 14. Rejection rate of the initial feature rejecter on the testing samples. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

where $\mathcal{N}(x|\mu,\Sigma)=\frac{1}{\sqrt{(2\pi)^g\cdot|\Sigma|}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$ is a multi-variant Gaussian function, while $\lambda^{[i]}=\{\omega_g^{[i]},\mu_g^{[i]},\Sigma_g^{[i]}\}_1^G$ is the set of parameters of person i with G Gaussians.

Its parameters are optimized by the Expectation Maximization (EM) algorithm. Due to the vectors being treated as i.i.d, information about the topology of the face is in effect lost. While at first this may seen counter-productive, the loss of topology in conjunction with overlapping blocks provides a useful characteristic: the precise location of face areas is no longer required namely being robust to imperfect face detection as well as a certain amount of in-plane and out-of-plane rotations.

For optimization by Expectation Maximization (EM), a fixed number of Gaussians should be set to describe those faces. As the matter of fact, the number of Gaussians affects the accuracy of the face model significantly. More Gaussians can give more precise face model, but it may not converge due to the limited training data. In order to ensure the convergence of each face model, the smallest Gaussian number of the training faces is selected to initialize EM, which can be referred to as Fixed Gaussian Mixture Model (FGMM).

3.3. Adaptive Gaussian Mixture Model (AGMM)

Unlike the FGMM which uses a fixed number of Gaussians ($G=32$) to model the distributions of each face, we propose to use an adaptive number of Gaussian Mixture Model to represent each face. The number of Gaussians $G^{[i]}$ and the other parameters $\lambda^{[i]}=\{\omega_g^{[i]},\mu_g^{[i]},\Sigma_g^{[i]}\}_1^{G^{[i]}}$ are estimated from the training dataset by maximizing the Log likelihood (10) with iterative EM [57]:

$$\begin{aligned}\arg\max_{\lambda}\ln P(\mathcal{X}|\lambda^{[i]}) &= \arg\max_{\lambda}\sum_{n=1}^{N}\ln\{P(x_n|\lambda^{[i]})\}\\ &= \arg\max_{\lambda}\sum_{n=1}^{N}\ln\left\{\sum_{g=1}^{G}\omega_g^{[i]}\mathcal{N}(x_n|\mu_g^{[i]},\Sigma_g^{[i]})\right\}\end{aligned} \tag{10}$$

Fig. 9 shows the optimal number of Gaussians needed for the faces (64×80) in FERET dataset divided by 8×8 block with 4 overlapping pixels. According to the information given in the figure, we found that the minimum and maximum numbers of Gaussians for a face are six for the 50th face and twenty-eight for

the 4th face, respectively. For FGMM, if setting $G=6$ as the number of Gaussians for each face, the faces such as the 4th one which have large variations cannot be modelled well. Similarly, if using too many Gaussians like $G=28$, EM may not be able to converge in the 50th face, because the samples with high dimensions is too sparse to be used to build the face model with 28 Gaussians. However, this issue can be solved using AGMM as appropriate number of Gaussians can be obtained adaptively for each face, which can give on average a 5% gain in recognition on average.

To evaluate AGMM, some examples with misalignment on different scales and detection windows are shown in the top row of Fig. 10. It can be observed that the face images from the same person, even having misalignment problem, are more similar (high similarity probability) than those from different persons in the bottom row of Fig. 9.

4. Experimental verification

In this section, we present the experimental results of preprocessing, CHSD training and testing, body model learning, and the performance of the proposed face recognition algorithm on publicly available databases and our dataset to demonstrate the efficacy of our method.

4.1. Preprocessing

In surveillance, low-quality face images mainly result from motion blur or non-uniform colour, and false detection, which can be removed by a new High Frequency Threshold (HFT), colour normalization, and background information, respectively.

4.1.1. Colour normalization

To equalize the colour and remove camera noise, preprocessing is very important to improve the performance of face recognition. In our method, we incorporate preprocessing prior to feature extraction. First, we standardize all face images to 64×80 pixels, and then normalize them to similar colour scale. Instead of using histogram equalization, we build the colour model for each pixel as:

$$\overline{p}(x,y) = b + c \cdot p(x,y) \tag{11}$$

where $p(x,y)$ is the "uncorrupted" pixel. Removing the DC component only corrects for bias b. To achieve the robustness to the contrast variations, the set of pixels within each block are normalized to have zero mean and unit variance $\mathcal{N}(0,1)$, which can be calculated fast by integral image and integral square image used in Section 2.1. Some results are shown in Fig. 11.

4.1.2. Blurred image removing

For the fuzzy image, it contains relatively smaller amount of energy in high frequency than that of the sharp image. In HFT, the ratio between the high-frequency coefficients and the low-frequency coefficients of the face image which are defined as Fig. 12(a) is used as a threshold to remove the blurred image. Two examples and their corresponding ratios are illustrated in Fig. 12(b) and (c). But only the image with significant global blurring artefacts can be removed, the image with local blurring, like moving mouth, may not be detected by the HFT. However, the proposed OLPF can handle the case due to motion effect on face component. For the false detected face image, it can be filtered out by background mask and skin colour.

4.2. CHSD training and face detection

4.2.1. CHSD training

The training data consists of 3860 hand-labeled frontal HS, which are collected from datasets, such as INRIA [22], Caltech [24], ETHZ [58], SCFace [12], and our cameras network and internet. The samples cover varying lighting, different quality, age, gender, and pose. All data are cropped and scaled to 64×80 pixels. For negative samples, we collected 5000 images without human being including the natural images and texture images, which are cropped to form a total of 944,338,068 non-HS images. Some positive and negative examples are shown in Fig. 13.

Initial feature rejecter: In the initial feature rejecter, we trained 42 and 32 rejecters for the region variance and block difference, respectively, which can yield the rejection rate of 91.99% and detection rate of 100% on the testing dataset. The blue curve in Fig. 14 denotes the combined result of the two feature sets. As can be seen from the figure, the performance of initial feature rejecter approaches to be stable with increment of the number of features, so only 15 features consisting of variance features and difference features are selected to construct the initial feature rejecter. According to the information given in Fig. 14, we can find that with the first variance feature rejecter, about 52.29% of the non-HS images are removed while yielding 100% detection rate for the testing dataset.

Haar-like rejecter: In the proposed CHSD, the joint Haar-like feature is used, where the elementary feature block size of the Haar-like feature is 4×4. The parts of feature sets are listed in Fig. 4. The dataset used for training the Haar-like rejecter consists of 3860 positive samples and 31970 negative samples. Those samples, including the positive samples and negative samples, are input to the AdaBoost training system, and the features and the corresponding thresholds with the best performance of separating those samples are selected to construct a weak filter.

HoG classifier: For good performance suggested by [22,24] to extract the HoG feature, each detection window without smoothing ($\sigma = 0$) is divided into cells of size 8×8 pixels and each group of 2×2 is integrated into a block in a sliding fashion, and blocks overlap with each other by 50% vertically and horizontally. Each cell is mapped into a 9-bin Histograms of Oriented Gradient and each block contains a concatenated vector of HoG from all its cells. So a block is thus represented by a 36-dimensional feature vector that is normalized to a $\mathcal{L}2$-norm unit length. Each detection window with size of 64×80 is represented by 7×9 blocks, giving a total of 2268 features points per detection window. These features are then classified by a soft linear SVM provided by SVMLight.

A contribution of our work to object detection is the integration of Haar-like features and HoG features into a cascade framework, which equips the HS detector with strong rejection ability without accuracy loss. A cascade of classifiers is employed to reject as many non-HS samples as possible at the earliest stages, which can efficiently reduce

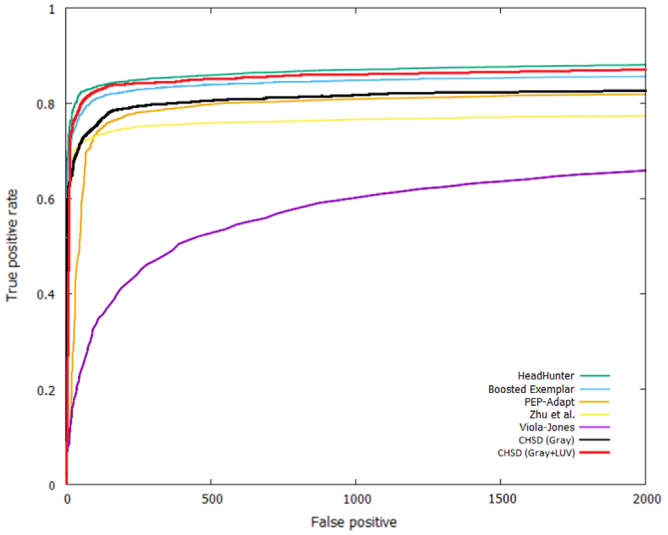


Fig. 15. Precision/recall curves of face detection methods on Pascal face dataset.

the detection time for the real-time system. The first rejecter, "*initial feature rejecter*", rejects almost 91.99% of the non-HS samples while retaining the detection rate of 100%. The second rejecter with 25 boosted Haar-like features can achieve 97.6% rejecter rate with 0.26% false positive rate. The following part is the SVM classifier which makes a final decision for the candidate regions.

4.2.2. Face detection on Pascal face dataset

The proposed face detection method has been evaluated and compared to state-of-the-art methods, including Mathias et al. [17], Boosted Exemplar [59], PEP-Adapter [60], Zhu and Ramanan [61], and Viola–Jones in OpenCV. We adopt the PASCAL VOC precision–recall protocol for object detection (requiring 50% overlap) and the results are shown in Fig. 15.

Apparently, our face detection method CHSD is competitive to state-of-the-art face detectors, such as HeadHunter and Boosted Exemplar, and outperforms the others on Pascal face dataset. Specifically, if using both Gray and LUV information, our proposed CHSD is the second best and only slightly worse than HeadHunter. It is worth noting that our CHSD is more efficient than Boosted Exemplar and HeadHunder as shown in Table 1, and can work in real-time applications.

4.2.3. Face detection on surveillance videos

We also tested the face detection in real surveillance videos, where the faces are often of low quality. So face detection becomes very challenging due to image blurring, compression noise, and variations in pose and illumination. In the testing, totally 2000 images were collected from real surveillance frames to build a challenging dataset with annotation for face detection validation. The whole performance (i.e., detection rate) is given in Table 2. Some examples are given in Fig. 16. Only the HeadHunter and Viola–Jones face detectors are included as their codes are released by the authors. Apparently, the proposed method is superior over HeadHunter in real detection tasks, which demonstrates the robustness of our detector (from Head–Shoulder to Face). It is worth noting that we have included some Head–Shoulder training samples with different viewpoints, e.g., $\pm 30^\circ$ in pan and $0^\circ - 60^\circ$ in tilt. So the proposed CHSD may also handle some side view detection.

The tradeoff between speedup and accuracy was investigated by two experiments: detecting time vs. the number of rejecters shown in Fig. 17 and accuracy vs. the number of rejecters (Fig. 18). In Fig. 17, we found that the first twenty rejecters can reject more than 90% non-HS region with detection time decreased to 56 ms. Adding more rejecters gains less computation time until the number of rejecters reaches 54. The detection time will increase when the number of rejecters is bigger than 54. This is because to classify the left HS and non-HS region samples, a more elegant rejecter is needed with more complicated features to be constructed. So, it also needs more time to do classification. The detection time is below 50 ms when the number of rejecters is between [30, 60] and the best number in terms of efficiency is 40. But adding more rejecters will degrade the performance of CHSD (accuracy decreased) due to high risk of making mistakes.

Table 1
Average detection time on Pascal VOC dataset.

Method	Time (ms)
HeadHunter	100
Boosted Exemplar	189.5
PEP-Adapt	>800
Zhu et al.	4000
CHSD (Gray)	13
CHSD (Gray+LUV)	34

Table 2
Comparison of face detection on surveillance frames.

Methods	Viola–Jones	HeadHunter	Our (Gray+LUV)
Detection rate	51.3%	78.6%	83.9%

Therefore, to achieve high efficiency (rejection rate) and accuracy (detection rate), we used 40 rejecters (15 at layer one and 25 at layer two) in the experiments.

4.3. Face recognition

In FGMM and AGMM, the face image is divided into blocks of 8×8 pixels with 4 overlapping pixels for extracting the OLPF feature. For a 64×80 face image, it results in 1073 feature vectors per face, and each feature vector contains 64 phase histogram bins (down-sampling the phase histogram bins to 64).

4.3.1. FERET dataset

For the FERET dataset, we selected nine poses (at -60°, -40°, -25°, -15°, 0°, $+15^\circ$, $+25^\circ$, $+40^\circ$, and $+60^\circ$), one illumination and one expression for each subject. In order to test the robustness to image blur, we added blurred images (with blurring kernels $\sigma = 1, 2, 3$), all together a total of 2758 images with 197 subjects. We use the frontal image (0°) as the gallery and others as the probe images. Tables 3 and 4 show the comparisons with existing methods on pose, illumination, expression, and blur variations. Clearly, AGMM has high recognition rate and outperforms the other algorithms except MDF. Because MDF generates a virtual image at the pose of the gallery image for the probe image through the 3D Morphable Displacement Field. And FGMM also comparable with state-of-art, such as StackFlow. In Table 4, we note that some algorithms are excluded, because they cannot handle the variations on illumination, expression and blur or not list the results in their papers.

4.3.2. Labeled Faces in the Wild

Labeled Faces in the Wild (LFW) [75] is an image dataset for unconstrained face recognition. It contains more than 13,000 face images collected from the web with large variations in pose, age, expression, illumination, etc. In our experiments, we followed the most restricted protocol [68], which splits the dataset into ten subsets with each subset containing 300 intra-class pairs and 300 inter-class pairs. The performances are measured by using 10-fold cross-validation.

We compare the proposed face recognition methods on LFW with the image-restricted protocol, and compared to the state-of-the-art methods such as [69,42,70–73]. The ROC curves of different methods are shown in Fig. 19, where the results of baselines are obtained from the official website of LFW.

Apparently, the proposed AGMM outperforms most methods except the PEP-based methods, such as Eigen-PEP and POP-PEP. However, PEP-based method normally takes the deep hierarchical architecture, and build representation as the concatenation of sequences of appearance descriptors (e.g., SIFT) with multiple layer fusion structure (coarse-to-fine). Therefore, such method is very computational expensive and cannot work at real-time speed. In contrast, our method is quite efficient and the recognition step only takes several milliseconds (<10 ms) to process one surveillance 4CIF frame (576×702) frame on a desktop with general dual-core 2.4 GHz CPU. The performance is competitive to PEP-based methods and very close to Eigen-PEP. The whole proposed recognition system can work in real-time speed (detection + recognition <44 ms per frame), and thus might be more promising for practical surveillance tasks.

Fig. 16. Face detection results on surveillance frames. First row: Viola–Jones' results; second row: HeadHunter's results; last row: our results.

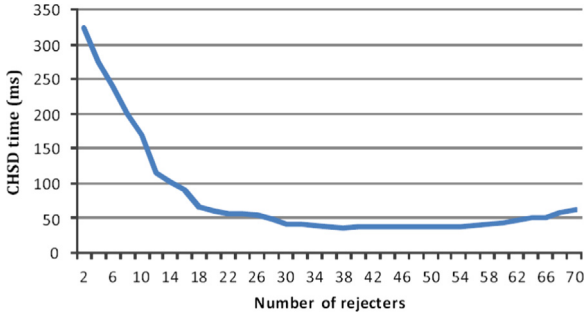


Fig. 17. CHSD detection time vs. number of rejecters.

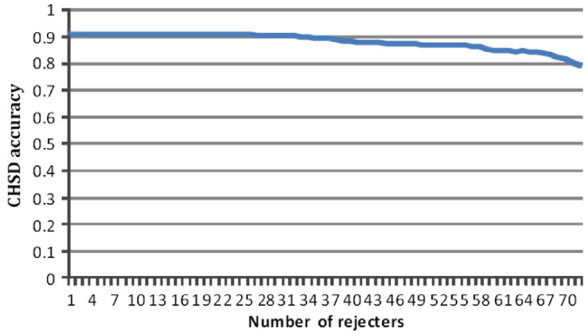


Fig. 18. CHSD detection accuracy vs. number of rejecters.

4.3.3. *Our dataset*

We also build a dataset with totally 9164 colour images collected from an indoor surveillance camera. Some samples are given in Fig. 2. For this dataset, the recognition rate of our method can reach 82.6% by using OLPF+FGMM and 84.9% by using OLPF+AGMM, respectively. Table 5 shows the recognition results for different descriptors (released source code) on our dataset. Both FGMM and AGMM with OLPF feature outperform other methods. It is noticed that [71] and FGMM have the similar performance in LFW and our dataset. The recognition rates for individuals are given in Fig. 20. For ID8 face images, it includes many images with severe expressions and noisy images which adversely affect the performance of our method.

Compared to PCA, we use the same number of training samples (six images) of each subject and the others for testing. The recognition rate of PCA is 46.3%. With the same training data, PCA is much worse than our method, because PCA needs accurate alignment and is sensitive to variations of pose, illumination, and blurring. While as illustrated in Fig. 21, the size of training set rarely affects the recognition rate of our proposed method. But the number of overlapping pixels significantly impacts on the

Table 3
Comparison with existing algorithms on FERET with pose variation (G=25 for FGMM).

Method	Pose							
	$-60°$	$-40°$	$-25°$	$-15°$	$+15°$	$+25°$	$+40°$	$+60°$
Eigenface [37]	3.2	8.5	23.7	54.3	49.7	36.1	11.5	5.2
MRPH [56]	NO	NO	85.6	88.2	88.1	66.8	NO	NO
FRR [55]	NO	NO	83.6	93.4	100	72.1	NO	NO
PLS [66]	39.6	59.3	76.5	76.8	77.3	72.9	53.8	37.9
StackFlow [65]	48.1	70.4	89.3	96.2	94.1	8.92	62.7	42.9
MDF [67]	87.5	97.2	99.4	99.7	100	99.4	98.1	92.0
FGMM	40.8	73.4	87.3	95.9	96.6	78.1	65.3	43.1
AGMM	56.4	80.6	91.3	100	100	88.4	76.8	58.1

recognition rate. That is because, in the high-dimensional space, overlapped samples can provide more complete clusters, which can be easily modelled by AGMM. In Table 5, although LPQ and LFD use the phase information, they cannot handle the pose variation and misalignment as good as our method due to the lack of a robust face model, like FGMM or AGMM.

It is noted that apart from pose variations, imperfect face localization [63] is also an annoying problem in a real life surveillance system. Imperfect localization results in translation as well as scale change, which adversely affects face recognition performance. The proposed face recognition method can solve the problem of imperfect localization, because our model is independent of the face topology. Some examples are illustrated in Fig. 10. In the first row, the imperfect face detection results in the face images with different locations and scales. In the AGMM model, the face images from the same person have higher similarity than those from different persons.

Table 4
Comparison with existing algorithms on FERET with illumination, expression and blur variation.

Algorithm	Accuracy	Expression	Blur (σ=1.0)	Blur (σ=2.0)	Blur (σ=3.0)
Eigenface [37]	58.0	36.8	78.9	64.7	53.4
LFD [64]	NO	NO	89.6	85.0	73.7
FGMM	81.3	75.8	99.6	93.5	81.6
AGMM	89.5	78.1	100	95.7	83.9

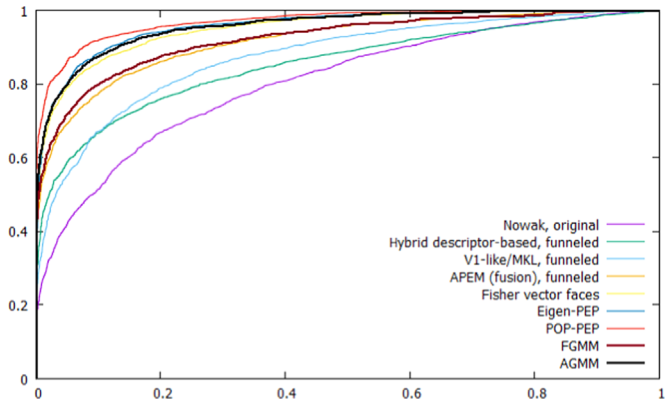


Fig. 19. Performance comparison on the restricted LFW.

5. Conclusion

A robust human detection and recognition system for surveillance is presented in this paper. The contributions can be summarized as follows: (1) we proposed CHSD with trained body model to solve the unconstrained face detection problem in surveillance; (2) we proposed a new face feature OLPF to represent the face discriminately which is not only invariant to blur but also robust to pose; (3) we proposed the FGMM and AGMM models to describe the distribution of the faces which are robust to both pose variation and imperfect detection; and (4) in preprocessing, we used the integral images to speed up the illumination normalization and removed blurred face images by HFT. Experimental results on FERET and real surveillance data show the superiority of

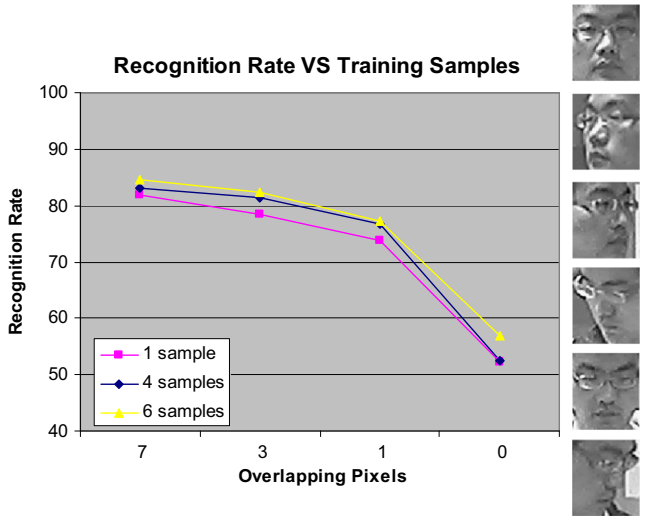


Fig. 21. Relationship between the size of training sample and recognition rate (AGMM); 1 sample and 4 samples correspond to the first one and four image in the right side image; and 6 samples use the entire samples listed in the right side.

Table 5
Comparison with existing algorithms on our dataset.

Method	PCA [37]	LBP [62]	LPQ [53]	LFD [58]	Fisher [71]	FGMM	AGMM
Recognition rate	46.3	49.1	57.9	69	82.4	82.6	84.9

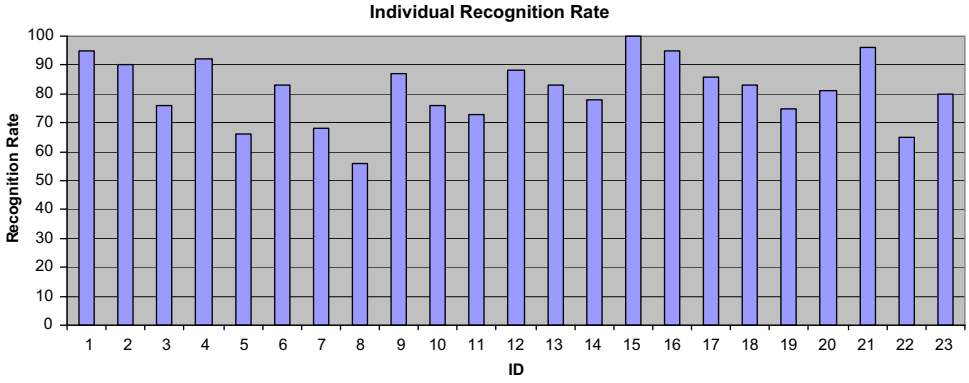


Fig. 20. Recognition rate of individuals in our dataset.

our proposed method over the existing algorithms. The human object detection and recognition scheme can be easily extended to implement on other interested objects with proper training dataset, like cars and animal detection and recognition.

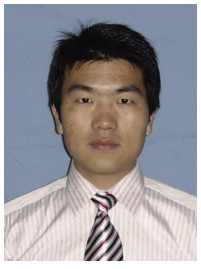
Acknowledgements

This work was supported by the NSFC Grant nos. 61203253, 61573222, and 61233014, Major Research Program of Shandong Province 2015ZDXX0801A02, Open Program of Jiangsu Key Laboratory of 3D Printing Equipment and Manufacturing 3DL201502, and Program of Key Lab of ICSP MOE China.

References

[1] M. McCahill, C. Norris, Urbaneye: CCTV in London, Centre of Criminology and Criminal Justice, University of Hull, UK, 2002.
[2] R. Gross, J. Yang, A. Waibel, Growing Gaussian mixture models for pose invariant face recognition, In: International Conference on Computer Vision, 2000, pp. 1088–1091.
[3] W. wang, R.P. Wang, Z.W. Huang, S.G. Shan, X.L. Chen, Discriminant analysis on Riemannian manifold of Gaussian distributions for face recognition with image sets, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2048–2057.
[4] D. Thomas, K.W. Bowyer, P.J. Flynn, Multi-factor approach to improving recognition performance in surveillance-quality video, In: 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, 2008, pp. 1–7.
[5] B. Chu, S. Romdhani, L. Chen, 3D-aided face recognition robust to expression and pose variations, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1907–1914.
[6] T. Hassner, S. Harel, E. Paz, R. Enbar, Effective face frontalization in unconstrained images, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4295–4304.
[7] B. Ma, A. Li, X. Chai, S. Shan, CovGa: a novel descriptor based on symmetry of regions for head pose estimation, Neurocomputing 143 (2014) 97–108.
[8] S. Kong, J. Heo, F. Boughorbel, Y. Zheng, B. Abidi, A. Koschan, M. Yi, M. Abidi, Multi-scale fusion of visible and thermal IR images for illumination-invariant face recognition, In: International Journal of Computer Vision, 2007, pp. 215–233.
[9] P.H. Hennings-Yeomans, S. Baker, B.V. Kumar, Simultaneous super-resolution and feature extraction for recognition of low resolution faces, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[10] Y. Jin, C. Bouganisr, Robust multi-image based blind face hallucination, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5252–5260.
[11] H. Zhang, J. Yang, Y. Zhang, N. Nasrabadi, T. Huang, Close the loop: joint blind image restoration and recognition with sparse representation prior, In: International Conference on Computer Vision, 2011, pp. 770–777.
[12] M. Grgic, K. Delac, S. Grgic, SCface—surveillance cameras face database, In: Multimedia Tools and Applications Journal, 2011, pp. 863–879.
[13] Y. Fang, J. Luo, C. Lou, Fusion of multi-directional rotation invariant uniform LBP features for face recognition, In: Third International Symposium on Intelligent Information Technology Application, 2009, pp. 332–335.
[14] H. Li, G. Hua, Z. Lin, J. Brandt, J.C Yang, Probabilistic elastic matching for pose variant face verification, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3499–3506.
[15] P. Viola, M. Jones, Robust real-time face detection, In: International Journal of Computer Vision, 2004, pp. 137–154.
[16] J. Chen, X. Chen, J. Yang, S. Shan, R. Wang, W. Gao, Optimization of a training set for more robust face detection, Pattern Recognit. 42 (11) (2009) 2828–2840.
[17] M. Mathias, R. Benenson, M Pedersoli, L.V. Gool, Face detection without bells and whistles, In: European Conference on Computer Vision, 2014, pp. 720–735.
[18] S. Rudrani, S. Das, Face recognition on low quality surveillance images, by compensating degradation, In: Image Analysis and Recognition, 2011, pp. 212–221.
[19] W. Zhao, R. Chellappa, A. Rosenfeld, P.J. Phillips, Face recognition: a literature survey, In: ACM Computing Surveys, 2003, pp. 399–458.
[20] Z. Cui, H. Chang, S. Shan, B. Ma, X. Chen, Joint sparse representation for video-based face recognition, Neurocomputing 135 (2014) 306–312.
[21] C. Zhang, Z.Zhang, A Survey of Recent Advances in Face Detection, Microsoft Research Technical Report, MSR-TR-2010-66, 2010.
[22] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005.
[23] M. Li, Z.X. Zhang, K.Q. Huang, T.N Tanl, Rapid and robust human detection and tracking based on omega-shape features, In: IEEE International Conference on Image Processing, 2009, pp. 2545–2548.
[24] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: an evaluation of the state of art, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 743–761.
[25] S. Baluja, H.A. Rowley, Boosting sex identification performance, In: International Journal of Computer Vision, 2007, pp. 111–119.
[26] R. Lienhart, J. Maydt, An extended set of Haar-like features for rapid object detection, In: Proceedings of the IEEE Conference Image Processing, 2002, pp. 900–903.
[27] S. Li, Z. Zhang, FloatBoost learning and statistical face detection, In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, pp. 1112–1123.
[28] Zhang et al. Informed Haar-like features improve pedestrian detection, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 947–954.
[29] T. Mita, T. Kaneko, O. Hori, Joint Haar-like features for face detection, In: Proceedings of the IEEE Computer Society Conference on Computer Vision, 2005, pp. 1619–1626.
[30] S. Zhang, R, Benenson, B. Schiele, Filtered channel features for pedestrian detection, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1751–1760.
[31] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. (1997) 119–139.
[32] P. Wang, Q. Ji, Learning discriminant features for multi-view face and eye detection, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 373–379.
[33] C. Liu, H.Y. Shum, Kullback-Leibler boosting, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2003, pp. 587–594.
[34] H. Li, K. Ngan, Q. Liu, FaceSeg: automatic face segmentation for real-time video, In: IEEE Transactions on Multimedia, U.S.A., 2009, pp. 77–88.
[35] P. Viola, M. Jones, D. Snow, Detecting pedestrians using patterns of motion and appearance, In: International Conference on Computer Vision, 2003, pp. 734–741.
[36] T. Joachims, Learning to classify text using support vector machines (Dissertation), Kluwer, 2002.
[37] M. Turk, A. Pentland, Face recognition using eigenfaces, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1991, pp. 586–591.
[38] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, pp. 711–721.
[39] X. He, S. Yan, Y. Hu, P. Niyogi, H.J. Zhang, Face recognition using Laplacianfaces, In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, pp. 328–340.
[40] J. Kim, J. Choi, J. Yi, M. Turk, Effective representation using ICA for face recognition robust to local distortion and partial occlusion, In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, pp. 1977–1981.
[41] X. Wang, X. Tang, Dual-space linear discriminant analysis for face recognition, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2004, pp. 564–569.
[42] L. Wolf, T. Hassner, Y. Taigman, Descriptor based methods in the wild, In: European Conference on Computer Vision, 2008.
[43] J. Ruiz-del-Solar, R. Verschae, M. Correa, Recognition of faces in unconstrained environments: a comparative study, In: EURASIP Journal on Advances in Signal Processing, 2009, pp. 1–20.
[44] X. Wang, X. Tang, Bayesian face recognition using Gabor features, In: Proceedings of ACM SIGMM 2003 Multimedia Biometrics Methods and Applications Workshop, Berkeley, CA, November 2003.
[45] T. Ojala, M. Pietikainen, T. Maenpaa, Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns, Lect. Notes Comput. Sci. (2000) 404–420.
[46] M. Bicego, A. Lagorio, E.Grosso, M. Tistarelli, On the use of SIFT features for face authentication, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2006.
[47] A. Albiol, D. Monzo, A. Martin, J. Sastre, A. Albiol, Face recognition using HOG-EBGM, Pattern Recognit. Lett. (2008) 1537–1543.
[48] P. Dreuw, P. Steingrube, H. Hanselmann, H. Ney, SURF Face: face recognition under viewpoint consistency constraints, In: British Machine Vision Conference, 2009.
[49] P. Hennings-Yeomans, S. Baker, B. Kumar, Simultaneous super-resolution and feature extraction for recognition of low-resolution faces, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[50] M. Gupta, S. Rajaram, N. Petrovic, T.S. Huang, Restoration and recognition in a loop, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 638–644.
[51] M. Nishiyama, H. Takeshima, J. Shotton, T. Kozakaya, O. Yamaguchi, Facial deblur inference to improve recognition of blurred faces, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1115–1122.
[52] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Ma Yi, Robust face recognition via sparse representation, In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, pp. 210–227.
[53] V. Ojansivu, J. Heikkilä, Blur insensitive texture classification using local phase quantization, In: Proceedings of the Image and Signal Processing, 2008, pp. 236–243.
[54] H. Li, G. Hua, Z. Lin, J. Brandt, J.C. Yang, Probabilistic elastic matching for pose variant face verification, In: 18th International Conference on Pattern Recognition, 2013, pp. 3499–3506.

[55] C. Sanderson, B.C. Lovell, Multi-region probabilistic histograms for robust and scalable identity inference, In: International Conference on Biometrics, 2009, pp. 199–208.
[56] T. Shan, B.C. Lovell, S. Chen, Face recognition robust to head pose from one sample image, In: 18th International Conference on Pattern Recognition, 2006, pp. 515–518.
[57] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. (1977) 1–38.
[58] A. Ess, B. Leibe, L.V. Gool, Depth and appearance for mobile scene analysis, In: International Conference on Computer Vision, 2007, pp. 1–8.
[59] H. Li, Z. Lin, J. Brandt, X. Shen, G. Hua, Efficient boosted exemplar-based face detection, In: Computer Vision and Pattern Recognition, 2014, pp. 1843–1850.
[60] H. Li, G. Hua, Z. Lin, J. Brandt, J. Yang, Probabilistic elastic part model for unsupervised face detector adaptation, In: IEEE International Conference on Computer Vision, 2013, pp. 793–800.
[61] X. Zhu, D. Ramanan, Face detection, pose estimation and landmark localization in the wild, In: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2879–2886.
[62] T. Ahonen, A. Hadid, M. Peitikaimen, Face description with local binary patterns: application to face recognition, In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, pp. 2037–2041.
[63] Y. Rodriguez, F. Cardinaux, S. Bengio, J. Mariéthoz, Measuring the performance of face localization systems, In: Image and Vision Computing, 2006, pp. 882–893.
[64] L. Zhen, T. Ahonen, M. Pietikainen, S.Z. Li, Local frequency descriptor for low-resolution face recognition, In: Automatic Face & Gesture Recognition and Workshops, 2011, pp. 161–166.
[65] A.B. Ashraf, S. Lucey, T. Chen, Learning patch correspondences for improved viewpoint invariant face recognition, In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.
[66] A. Sharma, D.W. Jacobs, Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch, In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 593–600.
[67] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, S. Shan, Morphable displacement field based image matching for face recognition across pose, In: European Conference on Computer Vision (ECCV), 2012, pp. 102–115.
[68] G. Huang, M. Mattar, T. Berg, E. Learned-Miller, et al., Labeled faces in the wild: a database for studying face recognition in unconstrained environments, In: Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, 2008.
[69] E. Nowak, F. Jurie, Learning visual similarity measures for comparing never seen objects, In: Computer Vision and Pattern Recognition (CVPR), 2007.
[70] N. Pinto, J.J. DiCarlo, D.D. Cox, How far can you get with a modern face recognition test set using only simple features? In: Computer Vision and Pattern Recognition (CVPR), 2009.
[71] K. Simonyan, O.M. Parkhi, A. Vedaldi, A. Zisserman, Fisher vector faces in the wild, In: British Machine Vision Conference (BMVC), 2013.
[72] H. Li, G. Hua, X. Shen, Z. Lin, J. Brandt, Eigen-PEP for video face recognition, In: Asian Conference on Computer Vision (ACCV), 2014.
[73] H. Li, G. Hua, Hierarchical-PEP Model for real-world face recognition, In: Computer Vision and Pattern Recognition (CVPR), 2015.
[74] W. Zhang, Y. Zhang, L. Ma, J. Guan, S. Gong, Multimodal learning for facial expression recognition, Pattern Recognit. 48 (10) (2015) 3191–3202.
[75] G. Huang, E. Miller, Labeled Faces in the Wild: Updates and New Reporting Procedures, Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.



Qiang Liu received the Ph.D. degree in Electronic Engineering from The Chinese University of Hong Kong in 2013. He received his M.S. degree from Huaqiao University in 2006, and B.E. degree from Anhui Institute of Architecture and Industry in 2004, respectively. His research interests include computer vision, image processing and pattern recognition.

Wei Zhang received the Ph.D. degree in Electronic Engineering from The Chinese University of Hong Kong in 2010. He is currently an associate professor of the School of Control Science and Engineering at Shandong University, China. His research interests include multimedia, computer vision, artificial intelligence, and robotics. He has published about 40 papers in prestigious journals and refereed conferences. He served as a program committee member and reviewer for various international conferences and journals in image processing, computer vision and robotics.

Hongliang Li received the Ph.D. degree in electronics and information engineering from Xi'an Jiaotong University, Xi'an, China, in 2005. From 2005 to 2006, he joined the Visual Signal Processing and Communication Laboratory, Chinese University of Hong Kong (CUHK), Shatin, Hong Kong, as a research associate. From 2006 to 2008, he was a post-doctoral fellow with the same laboratory in CUHK. He is currently a professor with the School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu, China. His current research interests include image segmentation, object detection and tracking, image and video coding, and multimedia communication systems.

King Ngi Ngan received the Ph.D. degree in electrical engineering from the Loughborough University, Loughborough, U.K. He is currently a chair professor with the Department of Electronic Engineering, Chinese University of Hong Kong. He was previously a full professor with the Nanyang Technological University, Singapore, and the University of Western Australia, Australia. He holds honorary and visiting professorships with numerous universities in China, Australia, and Southeast Asia. He is an associate editor of the Journal on Visual Communications and Image Representation, as well as an area editor of EURASIP Journal of Signal Processing: Image Communication and an associate editor for the Journal of Applied Signal Processing. He has published extensively including three authored books, five edited volumes, over 300 refereed technical papers, and edited nine special issues in journals. In addition, he holds 10 patents in the areas of image/video coding and communications. Ngan is a fellow of the IET and IEAust (Australia) and was an IEEE Distinguished Lecturer during 2006–2007. He served as an associate editor of the IEEE Transactions on Circuits and Systems for Video Technology. He chaired a number of prestigious international conferences on video signal processing and communications, and served on the advisory and technical committees of numerous professional organizations. He was a general co-chair of the IEEE International Conference on Image Processing (ICIP), Hong Kong, September 2010.