Bimodal Gender Recognition from Face and Fingerprint

Xiong Li[1*], Xu Zhao[1*], Yun Fu[2], and Yuncai Liu[1]
[1]Shanghai Jiao Tong University, Shanghai, 200240, China
[2]Department of CSE, University at Buffalo (SUNY), NY 14260, USA

Abstract

This paper focuses on multimodal gender recognition. To achieve a robust and discriminative performance for gender recognition, visual observations from both face and corresponding fingerprints are fused to serve for the task. The bag-of-words model is employed to structure the image representation. We propose a novel supervised method to construct the visual words, by which the redundant feature dimensions are discarded and the important dimensions for gender classification are highlighted. The dimension rearrangement is achieved by aligning the feature dimensions to a common normal vector of the hyperplane between categories. The Latent Dirichlet Allocation (LDA) model is extended to incorporate discriminative clues for supervised classification. We build the novel Discriminative LDA (D-LDA) model by maximizing the inter-class margins, which can significantly enhance the discriminative power of the whole model. Experiments on a large face and fingerprint database demonstrate the effectiveness of the proposed new feature and model. Complementary advantages benefited from face-fingerprint fusion to a robust gender recognition framework also get validated.

1. Introduction

As a basic capability of human beings, recognizing human gender plays an important role in our social activities. Imitating such ability in a computer is one of the critical tasks in Computer Vision and Pattern Recognition (CVPR) research. Potential applications widely appear in social interactions, surveillance, forensics, criminalistics, entertainment, marketing, and military affairs. Currently, machine based gender recognition is mainly using various kinds of human biometric features, such as face [28, 26, 18], fingerprint [25, 3], foot shape [27], teeth [20], and gait [13], *etc*.

In many existing works [16, 26, 28] in the CVPR field, the visual observation of human face is often selected as an important biometric cue for gender perception. Although human face provides discriminative information for gender recognition and is easier to capture than other biometric modalities, it is usually sensitive to the change of environmental conditions such as illumination, head pose, resolution, and so forth. Instead, fingerprint is more robust to these variations due to the well-controlled sensory input. In order to achieve a robust and discriminative performance for gender recognition, we propose to fuse visual information from both face and corresponding fingerprints.

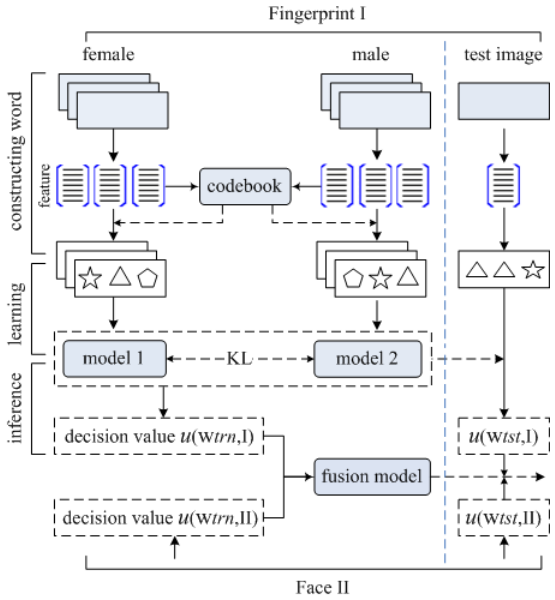


Figure 1: Flow chart of the proposed bimodal gender recognition framework through face and fingerprint fusion.

Previous works on fingerprint based gender recognition usually make use of specially designed features, such as ridge density, ridge count, ridge and valley thicknesses, finger size and white line count, *etc*. [3, 1, 15]. Extracting these features requires specially designed algorithm and relative high quality images which are expensive and tedious in practical applications. In this paper, we extend to use a more general and effective feature representation by introducing the *bag-of-words* model [8, 9], which is designed to build on Local Binary Pattern (LBP) [17] patches.

[*] indicates equal contribution.

978-1-4244-6985-7/10/$26.00 ©2010 IEEE

With human vision experience, it is highly probable that the gender mystery is encoded in the critical local regions of the human face and fingerprint, other than the whole. Patch based representation provides a well-suited mechanism to capture the salient parts of both face and fingerprint for gender recognition. Among the majority works to date, the bag-of-visual-words are usually constructed by unsupervised means, in which only the most common and frequent patterns in the entire training set are captured. Such representations may provide limited discriminative power in a targeted feature space for a specific classification task.

In this paper, we design a novel supervised method for visual word construction to strengthen the discriminative power of image representation. We start to construct the patch set from image grids. Each dimension in these patches plays a different role for gender classification. So we rearrange and then reselect these dimensions by aligning them to a common normal vector of the hyperplane between categories. This method greatly enhances the discriminative power of the bag-of-words representation.

In addition, this kind of representation can be naturally embedded into a generative framework, Latent Dirichlet Allocation (LDA) [5], for gender recognition purpose. In this context, gender is determined by the combination of different latent topics which are composed of a set of visual words. We train the generative models for both categories, male and female. By computing the likelihood of the two generative models, one can estimate the category label of a given test image. However, the normal LDA model works on each category separately without utilizing any inter-class information. Based on our previous work [14], we propose a novel model that explores the inter-class relationship to enhance the discriminative power of the LDA model. We achieve this objective by maximizing the inter-class margins, which is measured by KL divergence of the generative models describing male and female respectively. The novel model can also be expanded to other different applications besides gender recognition. We obtain the final recognition results by fusing the results from different modalities at the decision level.

Fig. 1 is a summary of our proposed gender recognition framework, which has been verified on a large face and fingerprint database recently collected in our group. Overall, the contributions of this paper are summarized as follows.

- We propose the first framework to estimate human gender by fusing both face and fingerprint information, which can achieve robust and discriminative feature representations for gender recognition.
- A novel supervised visual word construction method is presented. The formed visual words highlight the important dimensions while suppress those unimportant dimensions to adapt to a classification purpose.
- We propose a novel discriminative LDA model. By maximizing the inter-class margin while training the generative LDA model, the discriminative power of the whole model can be significantly enhanced.
- We collect a large database containing both face and corresponding fingerprint images to verify the performance of the proposed bimodal system.

The paper is organized as follows. We review the related work in Section 2. The supervised visual word construction method is introduced in Section 3. Section 4 describes the novel discriminative LDA model. Section 5 shows the experimental results. We draw the conclusions at last.

2. Related Work

Gender recognition has gradually attracted much attention from the CVPR field. While physiologists and crime-detection experts attack the gender estimation problem mainly through physiological features, visual information from human face is the mostly used visual observation for gender recognition. Many existing methods have been utilizing face images for gender recognition [4, 16, 26, 28, 23]. Usually these methods use global features like raw pixels, discriminative classification algorithms such as SVM [16], and Boosting [4, 28]. In our work, the local patch based image representation and the generative model are combined to capture subtle visual differences in gender patterns.

Fingerprint is another prevalent visual observation for gender recognition. Compared to face, fingerprint is more robust to the change of illumination and pose due to the special sensory structure. Therefore fingerprint based gender recognition is relatively simple and robust.

Besides face and fingerprint, other visual observations such as gaits [13, 29, 22], hand shape [2], foot shape [27], and human body [11], *etc*. are also employed in gender recognition. It is worth noting that fusing different modalities to improve the performance of gender recognition has been successfully validated [22, 29]. In [22], by fusing face and gait with canonical correlation analysis at the feature level, the performance of gender recognition obtains significant improvement. In [29], multi-view gait is introduced to fuse with face information for this task. SVM is used as classifier in [22, 29].

The works presented in [12, 19] are similar to ours in utilizing visual cues, where face and fingerprint information are fused to recognize human ID. However, in these works, face and fingerprint are not collected from the same subject and only the virtual subjects are used by assuming face and fingerprint are statistically independent for an individual [19]. Moreover, the features, classifiers, and fusion strategies are totally different from our framework. In this paper, we use a database that contains faces and corresponding ten fingerprints from the same subjects. To our best knowledge, this is the first database with such structure.

3. Image Representation

In this work, both face and fingerprint images are represented by the bag-of-words model. With human vision experiences, it is highly probable that some local regions of the human face play more important roles than other regions for gender recognition. This intuitive observation motivates us to represent the images within the patch based framework, which can provide the convenience to capture discriminative parts of the human face for gender recognition. The local patches are collected from the regularly placed image grids. Each patch is described by a LBP feature.

3.1. Supervised Visual Words Construction

The basic component of bag-of-words model is the visual words constructed from the local patch pool. However, the features extracted from these patches usually contain a lot of redundant dimensions noising the recognition process. Furthermore, normal patch based feature representation views each local patch as an independent part of the whole image but neglects each other's inner correlation. Better representation can be formed if the correlation between local features is considered on a much higher level for a special classification task. To this end, we develop a supervised approach [14] to construct the visual words.

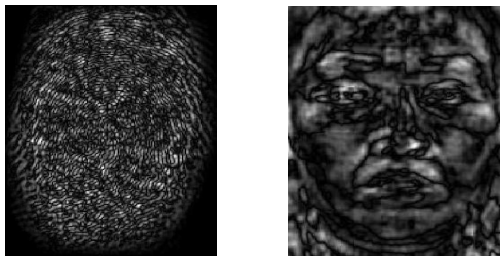
For an image I and its grid patch set $\{P_j\}_{j=1}^n$, we get the feature set $\{\mathbf{v}_j\}_{j=1}^n$ by extracting the local binary pattern on each grid patch, where $\mathbf{v}_j \in \Re^m$. Let $\mathbf{x}_i = (\mathbf{v}_1^T, \ldots, \mathbf{v}_n^T)^T$ denotes the feature of image I. Then one can get the normal vector $\vec{n} = (\mathbf{s}_1 y_1, \ldots, \mathbf{s}_t y_t)\hat{\alpha}$ of the decision hyperplane, between male and female categories from training samples, where support vector [24] $\mathbf{s}_i$ and $\hat{\alpha}$ can be obtained by maximizing the object function,

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j,$$

$$\text{s.t. } 0 \leq \alpha_i \leq \gamma, \ \sum_{i=1}^{N} \alpha_i y_i = 0,$$

where y_i is the label of $\mathbf{x}_i$ and N is the number of training samples. $\hat{\alpha}$ is composed of non-zero α_i. It is a constrained quadratic programming problem that can be solved by conventional methods. Each component of normal vector $\vec{n}$ actually measures the contribution of the corresponding dimension of $\mathbf{x}$ for classification. Dimensions with large values in $\vec{n}$ are preferred for feature reconstruction. Then, we rearrange the dimensions of $\mathbf{x}$ according to their weights $\vec{n}$ and get the new feature $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_{m \times n})^T$. Working on $\tilde{\mathbf{x}}$, we can reconstruct the word [1] set $\{\mathbf{w}_i\}_{i=1}^k$ for image I sequentially, where $\mathbf{w}_i = (\tilde{x}_{(i-1)\times l+1}, \ldots, \tilde{x}_{i \times l})^T$ is the i-th word with word length l. Note that usually the determination of k, l is subject to the constraint $k \times l \leq m \times n$.

[1] Although we call the reconstructed local features as *words* here, actually they are different from the visual words in the formed code book.



(a) (b)

Figure 2: Fingerprint and face images generated from normal vector of the decision hyperplane, which is built based on all the training samples. Brighter pixels are more salient for gender classification.

If l is small (namely the length of each word is short), inevitably the classification efficacy will get degraded. However, if the value of l is too large, less words will be available for the following visual words construction with a fixed training set. It is a tradeoff. To handle this situation, we turn to repeatedly select the training set randomly to construct words with relatively large l. Therefore the size of the word set gets augmented. One drawback in doing so is that the decision hyperplanes are varying with the change of the training set. It will degrade the performance of the formed word set. We solve this problem by aligning all the hyperplanes to a common hyperplane which has a normal vector $\vec{n}_0 = \frac{1}{\sqrt{m \times n}}(1, \ldots, 1)$, where $m \times n$ is the dimension of the hyperplane. Along with the aligning process, the training samples in each training subset are also transformed to preserve the relative geometric relationship with the corresponding hyperplane. Suppose by s random selections of the training set, we can get a set of normal vectors, $\{\vec{n}\}_{i=1}^s$, of the decision hyperplane. The transform matrix A_s for the s-th training set subjects to the following constrain,

$$A_s \vec{n} = \vec{n}_0.$$

After the word set is generated, the final image representation can be obtained by regular bag-of-words modeling [8] (see Fig. 1). In Fig. 2, we plot face and fingerprint figures to illustrate the results of feature selection. In the figure, each pixel takes the corresponding value in the normal vector $\vec{n}$ with the same dimension as the original image. Raw pixels are taken as image features for simplicity.

4. Discriminative Latent Dirichlet Allocation

The Latent Dirichlet Allocation (LDA) [5] is originally designed for corpus modeling. It is introduced to categorize natural scene images in [8]. However, even in supervised scenario, LDA only utilizes the intra-class information and

neglects the discriminative inter-class information. In this section, we propose a novel algorithm to enhance the discriminative power of LDA. Different from the maximum entropy criteria proposed in [30], we introduce a margin term to maximize the between-class interval.

4.1. Latent Dirichlet Allocation

For category c, LDA describes its stepwise generative process from $\pi, \mathbf{z}$ to $\mathbf{w}$ by a joint probability

$$p(\mathbf{w}, \mathbf{z}, \pi \,|\, \alpha, \beta) = p(\pi \,|\, \alpha) \prod_{n=1}^{N} p(z_n \,|\, \pi) p(w_n \,|\, z_n, \beta), \tag{1}$$

with

$$\begin{aligned} p(\pi \,|\, \alpha) &= \mathrm{Dir}(\pi \,|\, \alpha), \\ p(z_n \,|\, \pi) &= \mathrm{Mult}(z_n \,|\, \pi), \\ p(w_n \,|\, z_n, \beta) &= \prod_{i=1}^{K} p(w_n \,|\, \beta_i)^{\delta(z_n^i, 1)}, \end{aligned}$$

where π, $\mathbf{z}$ and $\mathbf{w}$ represent Dirchlet variable as well as the parameter of multinomial distribution, a set of topics and a set of words, respectively. In this application, words encode physiological properties, such as the type of mouth and so forth. The topics characterize the high level combinatorial properties. The n-th word w_n is in the form of a V-dimensional indicative vector, where $w_n^j = 1$ if the j-th term is selected. Similar form for topic z_n. 'Dir' denotes the Dirichlet distribution parameterized by $\alpha_{K\times 1}$. 'Mult' denotes the multinomial distribution parameterized by $\pi_{K\times 1}$ and $\beta_{K\times V}$, where $\beta_{ij} = p(w_j \,|\, z_i)$.

Eq. (1) gives the likelihood by integrating over median variables π and $\mathbf{z}$,

$$p(\mathbf{w} \,|\, \alpha, \beta) = \int_{\pi} p(\pi \,|\, \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \,|\, \pi) p(w_n \,|\, z_n, \beta) \right) d\pi.$$

Category label of a test image can be determined by the model after the image is encoded into visual words as introduced in Section 3.1. For supervised learning task, model $p(\mathbf{w} \,|\, \alpha_c, \beta_c)$ is trained for category c and an unknown image $\mathbf{w}$ is classified by picking $\arg\max_c p(\mathbf{w} \,|\, \alpha_c, \beta_c) p(c)$.

4.2. Discriminative Variational Inference

Variational inference. LDA [5] estimates models by maximizing the lower bound of log likelihood $\log p(\mathbf{w} \,|\, \alpha, \beta)$. To attack intractable likelihood $p(\mathbf{w} \,|\, \alpha, \beta)$, variational distributions $q(\pi \,|\, \gamma) \sim \mathrm{Dir}$ and $q(z_n \,|\, \phi) \sim \mathrm{Mult}$ are introduced to approximate $p(\pi \,|\, \alpha)$ and $p(\mathbf{z} \,|\, \pi)$, respectively. Then,

$$q(\pi, \mathbf{z} \,|\, \gamma, \phi) = q(\pi \,|\, \gamma) \prod_{n=1}^{N} q(z_n \,|\, \phi_n), \tag{2}$$

with which the lower bound of log likelihood can be formulated using Jensen's inequality,

$$\begin{aligned} \log p(\mathbf{w} \,|\, \alpha, \beta) &= \log \int_{\pi} \sum_{z} \frac{p(\pi, \mathbf{z}, \mathbf{w} \,|\, \alpha, \beta) q(\pi, \mathbf{z})}{q(\pi, \mathbf{z})} d\pi \\ &\geq E_q[\log p(\pi, \mathbf{z}, \mathbf{w} \,|\, \alpha, \beta) - \log q(\pi, \mathbf{z})]. \end{aligned}$$

Denote the lower bound of $\log p(\mathbf{w} \,|\, \alpha, \beta)$ as $L(\gamma, \phi, \alpha, \beta)$. It has been verified [5] that

$$\begin{aligned} \log p(\mathbf{w} \,|\, \alpha, \beta) &= L(\gamma, \phi, \alpha, \beta) + \\ &\quad D_e(q(\pi, \mathbf{z} \,|\, \gamma, \phi) \parallel p(\pi, \mathbf{z} \,|\, \mathbf{w}, \alpha, \beta)). \end{aligned}$$

An estimated model has a maximized lower bound L so that KL divergence $D_e \to 0$. We have

$$q(\pi, \mathbf{z} \,|\, \gamma, \phi) \xrightarrow{D} p(\pi, \mathbf{z} \,|\, \mathbf{w}, \alpha, \beta). \tag{3}$$

This approximation lays a foundation for the following work. The parameter estimation process of D-LDA can satisfy Eq. (3) through, 1) at initialization step, parameters are initialized with estimated LDA parameters, as shown in Fig. 3, 2) at estimation step, it can be satisfied by the object function Eq. (5). With Eq. (3) we have

$$\begin{aligned} p(\mathbf{w} \,|\, \alpha, \beta) &= p(\pi, \mathbf{z}, \mathbf{w} \,|\, \alpha, \beta) \,/\, p(\pi, \mathbf{z} \,|\, \mathbf{w}, \alpha, \beta) \\ &\approx q(\pi, \mathbf{z}, \mathbf{w} \,|\, \gamma, \phi, \beta) \,/\, q(\pi, \mathbf{z} \,|\, \gamma, \phi). \end{aligned}$$

Because $\pi \sim \mathrm{Dir}(\gamma)$ and $z_n \sim \mathrm{Mult}(\phi)$ are independent in variational approximation, we have

$$E_p\left[\log p(\mathbf{w} \,|\, \alpha, \beta)\right] \approx E_p[\log p(\mathbf{w} \,|\, \beta, \mathbf{z})]. \tag{4}$$

We further consider the total classification margin $\mathcal{M}$. As shown in the Appendix, total margin $\mathcal{M}$ can be maximized by maximizing $\hat{D}(p(\mathbf{w} \,|\, \Theta_1) \parallel p(\mathbf{w} \,|\, \Theta_2))$ on c_1 samples and $\hat{D}(p(\mathbf{w} \,|\, \Theta_2) \parallel p(\mathbf{w} \,|\, \Theta_1))$ on c_2 samples separately at the training step. Therefore the discriminative information, in the form of classification margin, can be introduced to our model by maximizing the first term in training c_1 model and the second term in training c_2 model. We simply denote current model parameters as α, β and parameters of its opposite model as $\tilde{\alpha}, \tilde{\beta}$. Considering Eq. (4), the margin term for current model can be formulated as

$$\begin{aligned} D(\beta, \tilde{\beta}) &= D(p(\mathbf{w} \,|\, \alpha, \beta) \parallel p(\mathbf{w} \,|\, \tilde{\alpha}, \tilde{\beta})) \\ &= E_p\left[\log p(\mathbf{w} \,|\, \alpha, \beta)\right] - E_p\left[\log p(\mathbf{w} \,|\, \tilde{\alpha}, \tilde{\beta})\right] \\ &\approx E_p\left[\log p(\mathbf{w} \,|\, \mathbf{z}, \beta)\right] - E_p\left[\log p(\mathbf{w} \,|\, \mathbf{z}, \tilde{\beta})\right] \\ &= \sum\nolimits_{n=1}^{N} \sum\nolimits_{i=1}^{K} \sum\nolimits_{j=1}^{V} z_n^i \beta_{ij} (\log \beta_{ij} - \log \tilde{\beta}_{ij}). \end{aligned}$$

The formulation reaches an important conclusion that the KL divergence between current model and its opposite model, equally the classification margin term, is determined by the last distribution $p(\mathbf{w} \,|\, \alpha, \beta, \mathbf{z})$.

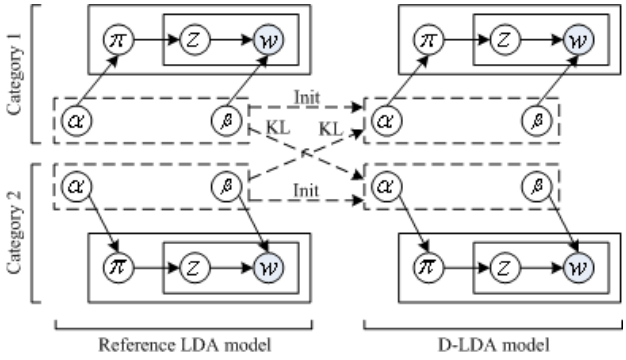


Figure 3: Graphical model representation of the D-LDA and reference models. Each D-LDA model is initialized by an estimated LDA model to satisfy Eq. (3), and the KL divergence namely the margin term is defined on both D-LDA model and LDA model instead of two LDA models.

We train the category model by maximizing the lower bound of log likelihood $L(\alpha, \beta, \gamma, \phi)$ and the margin term $D(\beta, \tilde{\beta})$ defined on the current model and its opposite model at the same time

$$\max_{\gamma,\phi,\alpha,\beta} L(\gamma, \phi, \alpha, \beta) + \lambda D(\beta, \tilde{\beta}), \tag{5}$$

where λ is a configuration parameter for $D(\beta, \tilde{\beta})$. We use well estimated LDA models as the opposite models on which the margin terms rely. Fig. 3 illustrates the graphical models of the proposed Discriminative LDA (D-LDA).

We employ variational method for inference and parameter estimation. Variational inference maximizes Eq. (5) with respect to parameters γ_i and ϕ_{ni} on the probability of an image, with constrain $\sum_{i=1}^{K} \phi_{ni} = 1$, which yields [5]

$$\begin{aligned} \phi_{ni} &= \frac{\beta_{iv} \exp\left(\Psi\left(\gamma_i\right) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right) - 1\right)}{\sum_{i=1}^{K} \beta_{iv} \exp\left(\Psi\left(\gamma_i\right) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right) - 1\right)}. \\ \gamma_i &= \alpha_i + \sum_{n=1}^{N} \phi_{ni}. \end{aligned}$$

Parameter estimation. The EM algorithm is employed for parameter estimation. The E-step is the variational inference described above. The M-step maximizes $\sum_m (L + \lambda D)$ with respect to parameters α_i and β_{ij} on the probability of all training images. α_i can be determined by

$$\begin{aligned} \frac{\partial \sum_m (L + \lambda D)}{\partial \alpha_i} &= M\left(\Psi\left(\sum_{j=1}^{K} \alpha_j\right) - \Psi(\alpha_i)\right) + \\ &\sum_{m=1}^{M} \left(\Psi(\gamma_{i_m}) - \Psi\left(\sum_{j=1}^{K} \gamma_{j_m}\right)\right). \end{aligned}$$

It is solved by Newton-Raphson algorithm after setting above derivative to zero and computing its Hessian matrix.

β_{ij} can be determined by considering a constrained optimization problem

$$\begin{aligned} \min_{\beta_{ij}} \quad & \sum\nolimits_m -(L + \lambda D) \\ \text{s.t.} \quad & \sum\nolimits_{j=1}^{V} \beta_{ij} = 1 \quad (i = 1, \cdots, K) \\ & 0 \le \beta_{ij} \le 1 \quad (i = 1, \cdots, K,\ j = 1, \cdots, V) \end{aligned} \tag{6}$$

The $K \times (V + 1)$ constrains can be divided into K groups by topic i. So the constrained optimization problem can be decomposed into K sub constrained optimization problems, each for a topic. These K constrained optimization problems can be effectively solved by Sequential Quadratic Programming (SQP) [6].

4.3. Decision Level Fusion

Face and fingerprint are two visual modalities in our framework. For each modality m, two models are trained for female c_1 and male c_2 respectively. Given the test person with word sets $\mathbf{w}_m$ for two modalities $m = 1, 2$, the likelihoods on the two models $p(\mathbf{w}_m \,|\, \alpha_m^c, \beta_m^c)$ can be computed from variational inference to determine the category label c. We define the decision value of modality m as

$$u_m = p(\mathbf{w}_m \,|\, \alpha_m^{c_1}, \beta_m^{c_1}) - p(\mathbf{w}_m \,|\, \alpha_m^{c_2}, \beta_m^{c_2}).$$

The fusion scheme integrates decision $\mathbf{u} = (u_1, u_2)^T$ of different modalities into a final decision u_0 with label set $\{c_1, c_2\}$. Let $\mathbf{h} = (h_1, h_2)^T$ denotes the threshold vector of $\mathbf{u}$. The Bayes risk of final decision [10] is

$$\mathcal{R} = \sum_{\mathbf{u}} \left(\sum_{i=1}^{2} \sum_{k=1}^{2} \alpha_{ik} P(u_0 = i \,|\, \mathbf{u}) P(\mathbf{u} \,|\, c_k; \mathbf{h}) P(c_k) \right),$$

where α_{ik} represents the decision loss, and $P(u_0 = i \,|\, \mathbf{u})$ can be well estimated from training results using nonparametric methods. In practice α_{ik} and $P(c_k)$ can be equally set for different k in this problem. We further write

$$P(\mathbf{u} \,|\, c_k; \mathbf{h}) = \sum_{i=1}^{2} P(\mathbf{u} \,|\, \mathbf{c}_i; \mathbf{h}) P(\mathbf{c}_i \,|\, c_k),$$

where $\mathbf{c}_i = (c_{1i}, c_{2i})^T$ and the m-th row of $C = (\mathbf{c}_1, \mathbf{c}_2)$ is the label set of modality m. $P(\mathbf{u} \,|\, \mathbf{c}_i; \mathbf{h})$ can be implemented as $P(u_m = j \,|\, C_{mi}; h_m)$ for $j = 1, 2$ and $m = 1, 2$ separately. $P(u_m = j \,|\, C_{mi}; h_m)$ and $P(\mathbf{c}_i \,|\, c_k)$ can be estimated from training results. Then the decision threshold vector $\mathbf{h}$ is determined by minimizing $\mathcal{R}$ with respect to $\mathbf{h}$.

5. Experiments

Experiments are conducted on a large data set [14], collected in our group, containing 197 females and 201 males

Figure 4: Sample images of two faces (female and male) [14] and the corresponding five fingerprints in our internal data set. Fingerprints from left to right correspond to thumb to litter finger.

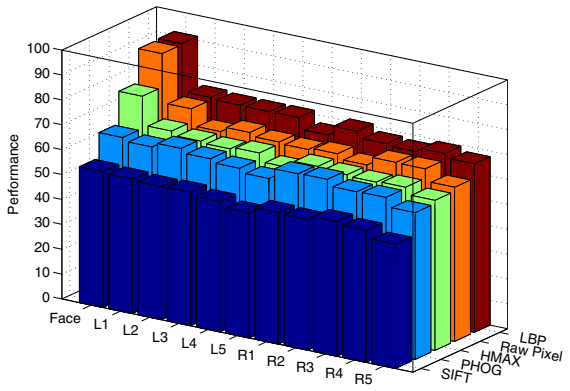


Figure 5: Performance of SIFT, PHOG, HMAX, Raw Pixel and LBP features on face and fingerprint modalities. L and R represent left and right hand respectively. The little finger is labeled as 1, and so on.

(Chinese) whose ages vary from 10 to 70. For each person, a 1280×1024 color bareheaded image and ten 328×356 gray fingerprint images are taken by digital camera and fingerprint sensor respectively. Fig. 4 shows some image samples of a female and a male in the data set [14]. In the experiments, 150 females and 150 males are selected, and the training samples and test samples are drawn randomly herein. Face and fingerprint images are normalized to 200×267 and 200×218 gray images respectively.

Firstly we conduct comparative experiments for both modality and feature selection. Five features–SIFT, PHOG [7], HMAX [21], Raw-pixel, and LBP–are evaluated on face and ten fingerprint modalities separately. As shown in Fig. 5, face and the left little finger (L1) outperform other modalities for most of the features, therefore they are selected as the experimental modalities. Similar results on fingerprint are also reported in [25]. As to features, LBP significantly outperforms other four features. We use LBP as the baseline feature hereafter.

For both fingerprint and face, LBP features are extracted

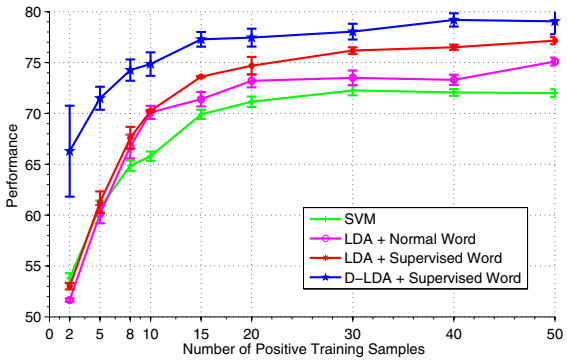


Figure 6: Performance comparison between SVM, LDA and D-LDA methods with raw feature, normal word and supervised word representations on the fingerprint modality.

on 12×12 grid patches. Before constructing supervised words, each feature is reduced to 20 dimensions from 59 by PCA. Then each face is represented as 36 words with 80 dimensions, while each fingerprint is represented as 36 words with 30 dimensions. For two modalities, both form code books with 110 coding centers, and share $\lambda = 0.001$, topic number $K = 3$.

To evaluate the performances of the models and features, we conducted the comparative experiments on four settings: 1) SVM for raw LBP feature without word construction, 2) LDA with normal word representation, 3) LDA with supervised words construction, and 4) D-LDA with supervised words construction. Following experiments share the same configuration that the number of negative training samples, positive testing samples and negative testing samples are fixed to 50, and the number of positive training samples is varying from 2 to 50. For each setting, experiment is repeatedly performed for 100 rounds and in each round the data is sampled randomly. The performance is measured by Correct Classification Ratio (CCR).

Fig. 6 shows the comparison results on the fingerprint modality. It can be seen that all four settings share a common trend that 15 positive training samples almost reach the peak-points. It indicates that the performance of fingerprint modality is hard to be improved through increasing the training samples. All word representation base methods outperform SVM, suggesting that local patches or words of fingerprint contain rich information to distinguish human gender. Furthermore, supervised words outperform normal words about 2% to 3% for more than 15 training samples. The advantages of D-LDA model also get verified. It outperforms other three settings at all scenarios. Especially for the case of small size training samples, it outperforms LDA with supervised words about 5% to 12%.

As a comparison, we explore three specially designed fingerprint features, Ridge Thickness to Valley Thickness

Table 1: Performance of specially designed fingerprint features, Ridge Thickness to Valley Thickness Ratio (RTVTR), Pattern Type Vector (PTR) and Left-right Hand Correspondence (LRC).

Feature	Dimensions	Performance (CCR)
RTVTR	1	61.23%
PTR	10	57.34%
LRC	5	55.76%

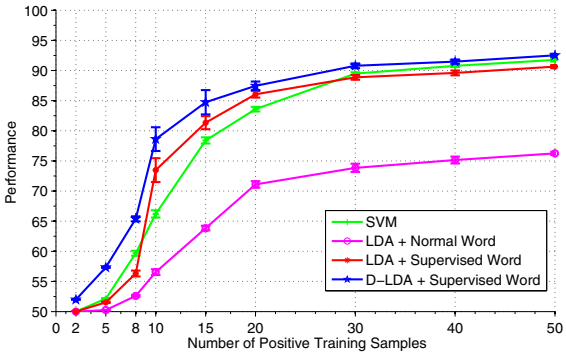


Figure 7: Performance comparison of SVM, LDA and D-LDA methods with raw feature, normal word and supervised word representations on the face modality.

Ratio (RTVTR), Pattern Type Vector (PTR) and Left-right Hand Correspondence (LRC) [3, 25] on our data set. LRC feature $\mathbf{x}$ is defined on a pair of corresponding fingerprints, such as the left thumb and right thumb, where $x_i = \text{Similarity}(L_i, R_i)$ for $i = 1, \cdots, 5$. These features are extracted from 160 females and 160 males. Table 1 shows their performances which are no more than 62%. The reason for the performance difference between [3, 25] and our experiments is that fingerprints in [3, 25] are collected by digital camera which are full and with high quality, while the quality of ours collected by fingerprint sensor is low.

The performance comparisons in face modality are shown in Fig. 7. It can be seen that the supervised word representation outperforms the normal word representation [8] about 15%. One reason to explain this is that the supervised word can catch more discriminative information than the normal word. Another reason is that the supervised word can catch global features such as face contour, but the normal word tends to miss them. We also find that with supervised word representation, both D-LDA and LDA have approximative performance with SVM under almost all scenarios. Compared to Fig. 6, the superiority of D-LDA over LDA is small, which suggests that the face modality benefits less from D-LDA than fingerprint modality.

A further experiment is conducted to validate our multimodal fusion scheme for gender estimation. Fig. 8 shows the performances of D-LDA and LDA on face, fingerprint

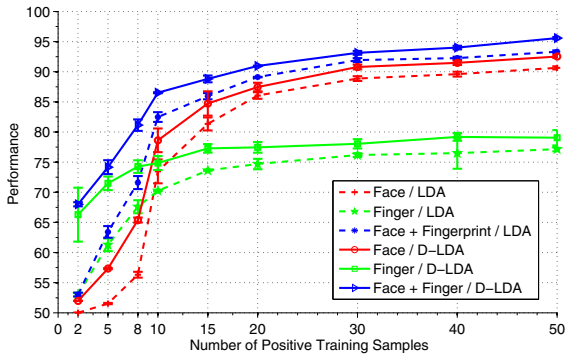


Figure 8: Performance evaluation of fingerprint and face modalities as well as their fusion. D-LDA and LDA algorithms with supervised word representation are employed.

modalities and their fusion with supervised word representation. Generally the face modality works well with more than 15 positive training samples. The fingerprint modality has good performance in few training samples scenario, which shows its complementarity to the face modality. By fusing clues from these two modalities, D-LDA with few training samples gives credible gender estimation, especially 8 samples reach a performance of more than 80%. It is meaningful as practical application usually works with few samples. Compared with LDA based fusion scheme, the D-LDA based fusion scheme significantly outperforms about 3% to 15%, especially for settings with few training samples.

We also compare the time cost of the normal and supervised word representations. The normal word representation usually forms a code book with 300 code centers from $272 \times N$ data points and codes a 200×218 image with 272 words, which takes 45 and 5 seconds respectively. The supervised word representation forms a code book with 110 code centers from $36 \times N$ data points and codes the same image with 36 words, which consumes 4 and 0.7 seconds respectively (the code book is formed by K-means algorithm and it follows nonlinear time cost along the number of data points). However, to estimate D-LDA parameters, LDA parameters have to be estimated beforehand, as shown in Fig. 3, which costs additional time. Moreover, solving Eq. 6 consumes more time than other steps, about 100 seconds for 50 training samples.

6. Conclusions

We proposed a novel framework for gender recognition. In this work, both face and fingerprint are firstly incorporated together to serve for gender recognition, in order to achieve a robust and discriminative performance. To capture the most salient parts in face and fingerprint for gender classification, the bag-of-words model is used to represent

the image. We proposed a novel supervised visual words construction scheme, by which the redundant feature dimensions are discarded and the important dimensions are highlighted. Such representation greatly enhanced the discriminate power of the whole algorithm. At the model level, to enhance the discriminative capability, we propose a discriminative LDA model, which explores the inter-class relationship. The efficacy of the proposed feature representation and the new model was verified by the comprehensive experiments conducted on our unique face and fingerprint database.

Acknowledgment

Thanks to China 973 program (2006CB303103), NSFC Key program (60833009) and national 863 program (2009AA01Z330) for funding.

Appendix: Classification Margin

For N images $\{\mathbf{w}_i\}_{i=1}^N$ of two categories whose models are parameterized by Θ_1, Θ_2, the discriminant function can be defined as

$$\mathcal{L}(\mathbf{w}_i \,|\, \Theta_1, \Theta_2) = \log \frac{p(\mathbf{w}_i \,|\, \Theta_1)}{p(\mathbf{w}_i \,|\, \Theta_2)}.$$

Then the total classification margin of all images can be formulated as

$$\begin{aligned} \mathcal{M} &= \sum_{i=1}^{N} y(\mathbf{w}_i)\mathcal{L}(\mathbf{w}_i \,|\, \Theta_1, \Theta_2) \\ &= \sum_{i=1}^{N_1} \log \frac{p(\mathbf{w}_i \,|\, \Theta_1)}{p(\mathbf{w}_i \,|\, \Theta_2)} + \sum_{i=1}^{N_2} \log \frac{p(\mathbf{w}_i \,|\, \Theta_2)}{p(\mathbf{w}_i \,|\, \Theta_1)} \\ &= \mathcal{M}_1 + \mathcal{M}_2, \end{aligned}$$

where $y(\mathbf{w}_i) = 1$ for $\mathbf{w}_i \in c_1$ and $y(\mathbf{w}_i) = -1$ for $\mathbf{w}_i \in c_2$ respectively. Note that $\mathcal{M}_1/N_1$ is actually the estimation of $E_{p(\mathbf{w} \,|\, \Theta_1)}[\mathcal{L}]$ on the samples of c_1. On the other hand, according to the definition of KL divergence, we have $E_{p(\mathbf{w} \,|\, \Theta_1)}[\mathcal{L}] = D(p(\mathbf{w} \,|\, \Theta_1) \parallel p(\mathbf{w} \,|\, \Theta_2))$. Therefore,

$$\begin{aligned} \mathcal{M}_1 &= N_1 \hat{E}_{p(\mathbf{w} \,|\, \Theta_1)}[\mathcal{L}] \\ &= N_1 \hat{D}(p(\mathbf{w} \,|\, \Theta_1) \parallel p(\mathbf{w} \,|\, \Theta_2)). \end{aligned}$$

Then the total classification margin can be expressed as the combination of two KL divergences,

$$\begin{aligned} \mathcal{M} &= N_1 \hat{E}_{p(\mathbf{w} \,|\, \Theta_1)}[\mathcal{L}] + N_2 \hat{E}_{p(\mathbf{w} \,|\, \Theta_2)}[-\mathcal{L}] \\ &= N_1 \hat{D}(p(\mathbf{w} \,|\, \Theta_1) \parallel p(\mathbf{w} \,|\, \Theta_2)) \\ &\quad + N_2 \hat{D}(p(\mathbf{w} \,|\, \Theta_2) \parallel p(\mathbf{w} \,|\, \Theta_1)). \end{aligned}$$

It suggests that the total classification margin $\mathcal{M}$ can be maximized by maximizing $\hat{D}(p(\mathbf{w} \,|\, \Theta_1) \parallel p(\mathbf{w} \,|\, \Theta_2))$ on c_1 samples and $\hat{D}(p(\mathbf{w} \,|\, \Theta_2) \parallel p(\mathbf{w} \,|\, \Theta_1))$ on c_2 samples separately at the training step.

References

[1] M. Acree. Is there a gender difference in fingerprint ridge density? *Forensic science international*, 102(1):35–44, 1999.

[2] G. Amayeh, G. Bebis, and M. Nicolescu. Gender classification from hand shape. In *CVPR Workshops*, 2008.

[3] A. Badawi, M. Mahfouz, R. Tadross, and R. Jantz. Fingerprint-based gender classification. In *The International Conference on IPCVPR*, 2006.

[4] S. Baluja and H. Rowley. Boosting sex identification performance. *IJCV*, 71(1):111–119, 2007.

[5] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] P. Boggs and J. Tolle. Sequential quadratic programming. *Acta numerica*, 4:1–51, 2008.

[7] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *ICCV*, 2007.

[8] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[9] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from Google's image search. In *ICCV*, 2005.

[10] A. Gunatilaka and B. Baertlein. Feature-level and decision-level fusion of noncoincidently sampled sensors for land mine detection. *IEEE Trans. on PAMI*, 23(6):577–589, 2001.

[11] G. Guo, G. Mu, and Y. Fu. Gender from Body: A Biologically-Inspired Approach with Manifold Learning. In *ACCV*, 2009.

[12] A. Jain, K. Nandakumar, X. Lu, and U. Park. Integrating faces, fingerprints, and soft biometric traits for user recognition. In *LNCS*, volume 3087, pages 259–269. Springer, 2004.

[13] L. Kozlowski and J. Cutting. Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics*, 21(6):575–580, 1977.

[14] X. Li, X. Zhao, H. Liu, Y. Fu, and Y. Liu. Multimodality gender estimation using bayesian hierarchical model. In *ICASSP*, 2010.

[15] D. LoESCH and N. Martin. Directional and absolute asymmetry of digital ridge counts. *Acta Anthropogenetica*, 6(2):85–98, 1982.

[16] B. Moghaddam and M. Yang. Learning gender with support faces. *IEEE Trans. on PAMI*, 24(5):707–711, 2002.

[17] T. Ojala, M. Pietikainen, and T. Maenpaa. Gray scale and rotation invariant texture classification with local binary patterns. In *LNCS*, volume 1842, pages 404–420. Springer, 2000.

[18] A. O'Toole, K. Deffenbacher, D. Valentin, K. McKee, D. Huff, and H. Abdi. The perception of face gender: The role of stimulus structure in recognition and classification. *Memory and cognition*, 26:146–160, 1998.

[19] A. Patra and S. Das. Enhancing decision combination of face and fingerprint by exploitation of individual classifier space: An approach to multi-modal biometry. *Pattern Recognition*, 41(7):2298–2308, 2008.

[20] G. Schwartz and M. Dean. Sexual dimorphism in modern human permanent teeth. *Am J Phys Anthropol*, 128(2):312–318, 2005.

[21] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *CVPR*, 2005.

[22] C. Shan, S. Gong, and P. McOwan. Fusing gait and face cues for human gender recognition. *Neurocomputing*, 71(10-12):1931–1938, 2008.

[23] M. Toews and T. Arbel. Detection, Localization, and Sex Classification of Faces from Arbitrary Viewpoints and under Occlusion. *IEEE Trans. on PAMI*, 31(9):1567–1581, 2009.

[24] V. Vapnik. *The nature of statistical learning theory*. Springer Verlag, 2000.

[25] J. Wang, C. Lin, Y. Chang, M. Nagurka, C. Yen, and C. Yeh. Gender determination using fingertip features. *Internet Journal of Medical Update*, 3(2), 2008.

[26] L. Wiskott, J. Fellous, N. Kruger, and C. von der Malsburg. Face recognition and gender determination. In *IEEE International Workshop on Automatic Face and Gesture Recognition*, 1995.

[27] R. Wunderlich and P. Cavanagh. Gender differences in adult foot shape:implications for shoe design. *Med Sci Sports Exerc*, 33(4):605–615, 2001.

[28] X. Xu and T. S. Huang. SODA-Boosting and its application to gender recognition. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, volume 4778, pages 193–204, 2007.

[29] D. Zhang and Y. Wang. Gender Recognition Based on Fusion of Face and Multi-view Gait. In *LNCS*, volume 5558, pages 1010–1018. Springer, 2009.

[30] J. Zhu, A. Ahmed, and E. Xing. Maximum Margin Supervised Topic Models for Regression and Classification. In *ICML*, 2009.