MULTIMODALITY GENDER ESTIMATION USING BAYESIAN HIERARCHICAL MODEL

Xiong Li [*], *Xu Zhao* [*], *Huanxi Liu* [*], *Yun Fu* [†] *and Yuncai Liu* [*]

[*]Institute of Image Processing & Pattern Recognition
Shanghai Jiao Tong University, 200240, Shanghai, China
{lixiong,zhaoxu,jadbm,whomliu}@sjtu.edu.cn
[†]Department of CSE, University at Buffalo (SUNY), NY 14260, USA
raymondyunfu@gmail.com

ABSTRACT

We propose to estimate human gender from corresponding fingerprint and face information with the Bayesian hierarchical model. Different from previous works on fingerprint based gender estimation with specially designed features, our method extends to use general local image features. Furthermore, a novel word representation called latent word is designed to work with the Bayesian hierarchical model. The feature representation is embedded to our multimodality model, within which the information from fingerprint and face is fused at the decision level for gender estimation. Experiments on our internal database show the promising performance.

Index Terms— Gender estimation, fingerprint and face, Bayesian hierarchical model, latent word representation, multimodality

1. INTRODUCTION

Gender estimation has been extensively studied, based on various kinds of human biometric features [1], such as face [2, 3], body [4], fingerprint [5, 6], hand shape [7], foot shape [8], and teeth [9], etc. Researchers in computer vision field usually seek gender hints from human face while physiologists and crime experts mainly tackle gender estimation problem through physiological features.

In this paper, we estimate human gender from both face and fingerprint [10, 11]. For fingerprint based gender estimation, most pervious works use the specially designed features, such as ridge count, finger size and white line count. However, these features require relative high quality images, which are hard to get in practical scenarios. We present a novel algorithm to deal with this problem in this paper. Unlike previous methods which use specialized features, our algorithm introduces local binary pattern [12], and then represents the candidate image with bag-of-words. We also design a novel word representation to avoid the shortages occurred in the normal bag-of-words model [13].

Thanks to China National 973 Program 2006CB303103, China NSFC Key Program 60833009 and National 863 program 2009AA01Z330 for funding.

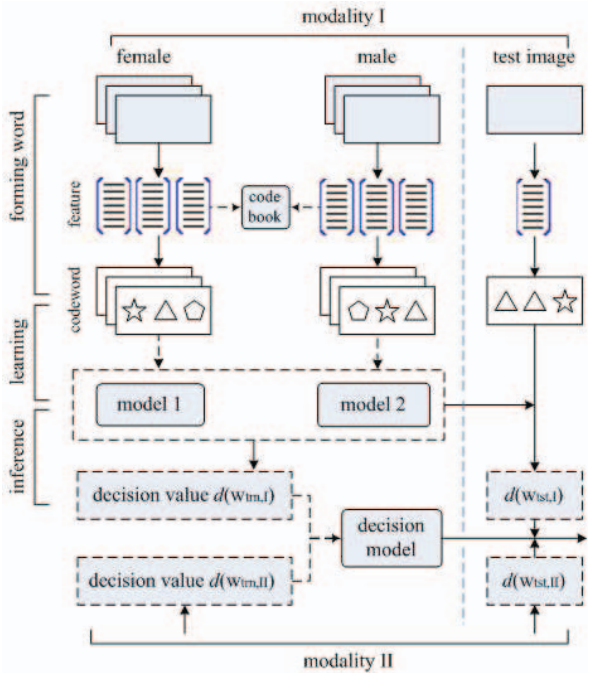

Fig. 1. Illustration chart of the algorithm.

We train the generative models within the framework of Bayesian hierarchical model, for both female and male categories. The gender of a test image then could be estimated by figuring out the likelihood of two generative models. The algorithm not only adapts to estimate human gender from both fingerprint and face, but also provides a general way to fuse multiple modalities for gender estimation. Fig.1 illustrates the flow of the algorithm. In sum, our contributions can be summarized as follows.

1. Estimate human gender by fusing both face and fingerprint information.
2. Design the multimodality fusion framework at the de-

cision level using the Bayesian hierarchical model.
3. Present a novel word representation to improve the efficacy of the normal bag-of-words model.

2. THE NOVEL WORD REPRESENTATION

Normal word representation [14][13] extracts local features on image grids, each feature giving a word. The word representation assumes that each grid is an integrated part; whereas cues from different grids actually form a better representation for their inner correlations. Therefore the combinations of several local features may generate more discriminative words. Inspired by this idea, we develop a novel word representation, which is called *latent word* in this paper.

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 $\mathbf{v}_j \in \mathbb{R}^m$. Let $\mathbf{x}_i = (\mathbf{v}_1^T, \ldots, \mathbf{v}_n^T)^T$ denote the feature set of image I. Then one can get the normal vector $\vec{n} = (\mathbf{s}_1, \ldots, \mathbf{s}_t)\hat{\alpha}$ of the decision hyperplane, between male and female categories from training samples, where $\hat{\alpha}$ is the non-zero element set of α. Support vectors [15] $\mathbf{s}_i$ and coefficients vector α can be determined by optimizing 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 sample $\mathbf{x}_i$. 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 word construction. 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 re-construct the word set $\{w_i\}_{i=1}^k$ for image I sequentially, where the ith word $w_i = (\tilde{x}_{(i-1)\times l+1}, \ldots, \tilde{x}_{i \times l})^T$ in which k is the word length constrained by $l \times k \leq m \times n$.Note that a word is defined as a local feature in the normal bag-of-words model, but a set of redefined components of the global feature in our representation.

Short words generally associate with weak decision ability, while short representations usually associate with the performance with large variance. In order to avoid the two problems simultaneously, we turn to select a sub training set randomly to construct words every time, and repeat the process roundly. These decision hyperplanes of different sub training sets are varying, leading that the word performance will degrade as combining words of different sub training sets. We attack the problem by transforming samples so that each hyperplanes is correspondingly rotated to a median hyperplane with normal vector $\vec{n}_0 = \frac{1}{\sqrt{m \times n}}(1, \ldots, 1)$, where $m \times n$ is the dimension number of the median hyperplane. The rotation matrix A is a solution of $A\vec{n} = \vec{n}_0$.

3. BAYESIAN HIERARCHICAL MODEL FOR MULTIMODALITY GENDER ESTIMATION

Previous works on Bayesian hierarchical model are mainly used in text topic modeling. In [16], latent Dirichlet allocation is designed for both supervised and unsupervised topic modeling. In [13], it is firstly applied in natural scene categorization. Inspired by the previous works, we introduce an effective modified version of the Bayesian hierarchical model for multimodality gender estimation. Generally the joint probability modeling the relationship between image words and its category can be formulated as

$$p(\mathbf{w}, \mathbf{z}, \pi, c|\alpha, \beta, \theta) = p(\mathbf{w}, \mathbf{z}, \pi|\alpha, \beta, c)p(c|\theta), \quad (1)$$

where $\mathbf{w}$, $\mathbf{z}$, π and c represent a set of image words, a set of word themes, a theme mixture, and a category respectively. Because we consider only two categories, female and male, the learning procedure can be simplified by separately modeling the two categories $p(\mathbf{w}, \mathbf{z}, \pi|\alpha_c, \beta_c), c = 1, 2$ for $p(\mathbf{w}, \mathbf{z}, \pi|\alpha, \beta, c)$. For category c, its Bayesian hierarchical model is a joint probability

$$p(\mathbf{w}, \mathbf{z}, \pi|\alpha_c, \beta_c) = p(\pi|\alpha_c) \prod_{n=1}^{N} p(z_n|\pi)p(w_n|z_n, \beta_c) \quad (2)$$

with

$$p(\pi|\alpha_c) = \text{Dir}(\pi_c|\alpha_c), \quad (3)$$

$$p(z_n|\pi) = \text{Mult}(z_n|\pi), \quad (4)$$

$$p(w_n|z_n, \beta_c) = \prod_{k=1}^{K} p(x_n|\beta_{ck})^{\delta(z_n^k, 1)}, \quad (5)$$

where Dir denotes the Dirichlet distribution parameterized by K-dimensional parameters α_c. Mult represents the multinomial distribution. β_c is a distribution parameter matrix of size $K \times T$, where K and T denote the total number of themes and code centers in the codebook respectively.

By integrating over the median variables π and $\mathbf{z}$, Eq.(2) gives the likelihood

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

Except for estimating $p(c|\theta)$ beforehand, the learning procedure of the model is similar to [16] by maximizing the log likelihood $\log p(\mathbf{w}|\alpha_c, \beta_c)$.

For a modality m where we denote fingerprint and face modalities as $m = 1, 2$ respectively, female and male models are learned. Given an unknown person with word set w_m, the likelihoods $p(w_m, c|\alpha_m, \beta_m, \theta_m) \propto p(c|w_m, \alpha_m, \beta_m, \theta_m)$ can be computed from Eq.(1)-(5) for determining the gender. We further define the decision value

$$d_m = p(w_m, c=1|\alpha_m, \beta_m) - p(w_m, c=2|\alpha_m, \beta_m).$$

Then the gender of modality m can be estimated by

$$c_m = \begin{cases} 1 & : \quad d_m \geq 0 \\ 2 & : \quad else \end{cases} .$$

The fusion model for two modalities is built at the decision level by modeling the conditional probability $p(c = 1|d_1, d_2)$. Instead of modeling it directly, we turn to model two simple probabilities

$$p(c=1|d_1, d_2) = \frac{p(c=1, d_1|d_2)}{p(d_1|d_2)}, \tag{6}$$

where distributions $p(c=1, d_1|d_2)$ and $p(d_1|d_2)$ can be well estimated by nonparametric methods using results from round test experiments.

4. EXPERIMENTS

Experiments are conducted on our internal database containing 197 females and 201 males of the Han nationality, whose age varies from 10 to 70. For each person, a 1280×1024 RGB bareheaded image and ten 328×356 grayscale fingerprint images are taken by digital camera and fingerprint sensor respectively. Fig. 2 shows some samples in our dataset. Beside face, the left little finger is selected as the experimental modality because it outperforms other fingers under all settings [5]. In the following experiments, 100 females and 100 males are selected and the training samples and the test samples are drawn randomly herein. Face and fingerprint images are normalized to 200×267 and 200×218 grayscale images respectively.

Fig. 2. Sample images of face and the corresponding five fingerprints in our data set.

For both finger and face modalities, local binary patterns are extracted on 12×12 grid patches. Before constructing latent words, each feature is reduced to 20 dimensions from 59 by PCA. In following experiments, the number of positive training samples is varying, and the number of negative training samples, positive testing samples, and negative testing samples are fixed at 50 respectively. We employ linear SVM classifier, that works on image feature $\mathbf{x}_i = (\mathbf{v}_1^T, \ldots, \mathbf{v}_n^T)^T$, as Section 2 mentioned, for comparative experiments.

The performance comparison on the fingerprint modality is shown in Fig. 3. All three methods follow a similar trend that 10 positive training samples almost reach the peak points, which means that the performance of fingerprint modality is hard to be improved through increasing the training samples. Both normal and latent word representations outperform SVM, suggesting that local patches or words of fingerprints contain rich information to distinguish the human gender. Fig. 3 also shows the advances of the latent word representation and the Bayesian hierarchical model because the method outperforms other two methods about 2% to 12% at all settings.

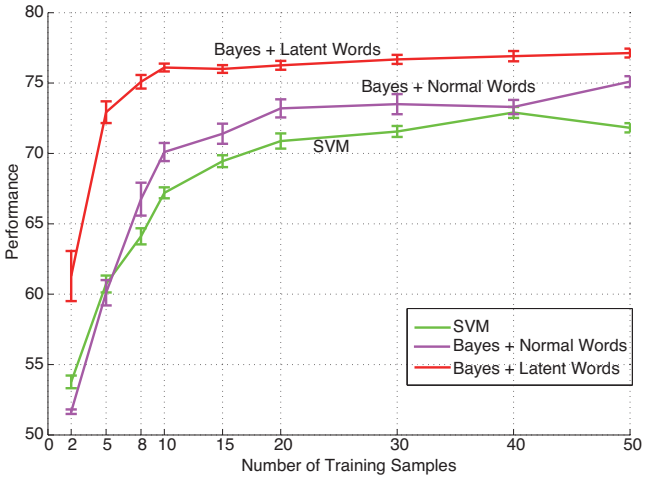

Fig. 3. Comparison of SVM and the Bayesian hierarchical model with both normal and latent word representations on the fingerprint modality.

As Fig. 4 shows, the performance of normal word representation [13] is lower than other two methods about 15%. A possible reason for it is that normal word representation tends to lose global information such as the face contour. We also find that the latent word representation with few training samples works well and outperforms SVM under almost all settings. Compared to Fig. 3, the performance difference between latent word representation and SVM is very small, which suggests that word representation is more adaptive for fingerprint than face.

A further experiment is conducted to validate our multimodality fusion model for gender estimation. The fusion model, as Eq. (6) shows, has to estimate two experiential distributions beforehand. Commonly the estimation samples are produced by round test beforehand on training set while incremental learning is also a feasible and effective scheme for the fusion model. Fig. 5 shows the performances of face and fingerprint modalities under the configuration of Bayesian hierarchical model and latent word representation. Generally the face modality works well with more than 15 positive training samples, and the fingerprint modality shows its complementarity to the face modality. By fusing cues from two modalities, few training samples could give credible gender estimation, especially 8 samples reach a performance of more than

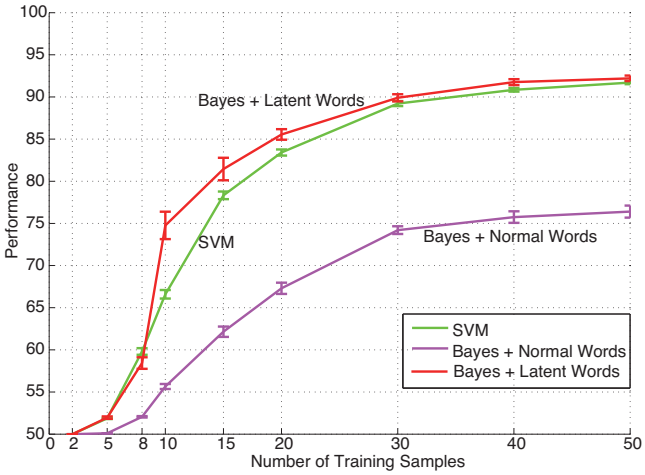

Fig. 4. Comparison of SVM and the Bayesian hierarchical model with both normal and latent word representations on the face modality.

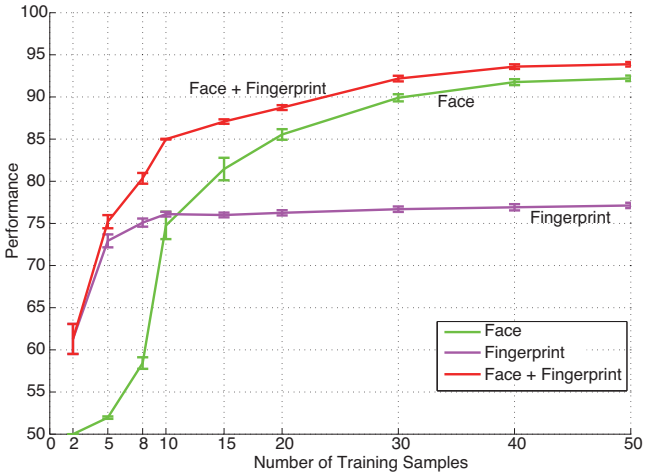

Fig. 5. Performance evaluation of the fingerprint and face modalities as well as their fusion. The Bayesian hierarchical model with latent word representation is employed.

80%. It is meaningful as practical application usually works with few samples.

We also compare the time cost of the normal and latent word representations. The normal word representation generally forms a codebook of 300 code centers from 272·N data points and codes a 200×218 image with 272 words, which take 45 and 5 seconds on a normal computer respectively. The latent word representation forms a codebook of 100 code centers from 30·N data points and codes the same image with 30 words, which take 4 and 0.7 seconds (k-means follows nonlinear time consuming along the number of data points) respectively. The whole computing time of our algorithm is only about $1/15$ of the pervious scheme [14][13].

5. CONCLUSIONS

We proposed a Bayesian hierarchical model for gender estimation. In the scheme, model is trained for each category. To better fit the model, each input image is represented as a bag-of-words, which extends previous methods of fingerprint estimation [5, 6] at the feature level. As a complementary, we also present a novel word representation called latent word to work with the Bayesian hierarchical model. It produces a more effective representation with much less words than the normal one. We introduce a probability model which works on multimodality fusion of fingerprint and face at the decision level. Experiments demonstrate that gender estimation benefits from both latent word representation and Bayesian hierarchical model. It also shows that the fusion of fingerprint and face information can achieve more robust and accurate performance for gender estimation than single modalities.

References

[1] A.K. Jain, K. Nandakumar, X. Lu, and U. Park, "Integrating faces, fingerprints, and soft biometric traits for user recognition," in *Proc.of Biometric Authentication Workshop, LNCS 3087*, 2004, pp. 259–269.

[2] X. Xu and T. S. Huang, "SODA-Boosting and its application to gender recognition," in *IEEE Int'l Workshop on AMFG*, 2007, pp. 803–806.

[3] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition and gender determination," in *Int'l Workshop on Automatic Face and Gesture Recognition*, 1995, pp. 92–97.

[4] G. Guo, G. Mu, and Y. Fu, "Gender from body: A biologically-inspired approach with manifold learning," in *ACCV*, 2009.

[5] J. Wang, C. Lin, Y. Chang, M. Nagurka, C. Yen, and C. Yeh, "Gender determination using fingertip features," *Internet Journal of Medical Update*, vol. 3, no. 2, 2008.

[6] A. Badawi, M. Mahfouz, R. Tadross, and R. Jantz, "Fingerprint-based gender classification," in *Int'l Conf. on IPCV*, 2006.

[7] G. Amayeh, G. Bebis, and M. Nicolescu, "Gender classification from hand shape," in *IEEE CVPR Workshops*, 2008.

[8] R. Wunderlich and P. Cavanagh, "Gender differences in adult foot shape:implications for shoe design," *Medicine and Science in Sports and Exercise*, vol. 33, no. 4, pp. 605–615, 2001.

[9] G. Schwartz and M. Dean, "Sexual dimorphism in modern human permanent teeth," *American Journal of Physical Anthropology*, vol. 128, no. 2, pp. 312–317, 2005.

[10] G.A. Khuwaja, "Merging face and finger images for human identification," *Pattern Analysis and Applications*, vol. 8, no. 1-2, pp. 188–198, 2005.

[11] 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*, vol. 41, no. 7, pp. 2298–2308, 2008.

[12] T. Ojala, M. Pietikainen, and T. Maenpaa, "Gray scale and rotation invariant texture classification with local binary patterns," in *Lecture Notes in Computer Science*. 2000, vol. 1842, pp. 404–420, Springer.

[13] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *IEEE CVPR*, 2005, vol. 2.

[14] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," in *IEEE ICCV*, 2005, vol. 2.

[15] V.N. Vapnik, "The nature of statistical learning theory," 2000.

[16] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.