How Fashion Talks: Clothing-Region-Based Gender Recognition

Shengnan Cai[1], Jingdong Wang[2], and Long Quan[1]

[1] Hong Kong University of Science and Technology, Hong Kong
scaiad@ust.hk, quan@cse.ust.hk
[2] Microsoft Research Asia, China
jingdw@microsoft.com

Abstract. In this paper, we investigate the gender recognition problem of people in photos via clothing information other than faces in the case of insufficient face specification. Similar to human's intuition on telling a person's gender from his/her dressing, we formulate this problem as a binary classification problem based on features extracted from semantic regions of clothing. Given a query image, we first apply category-level clothing parsing to divide the clothes into several semantic regions, such as blazers, shirts, jeans and so on. From each region, we obtain a local estimation on gender by classifying features describing color, texture and shape as middle level attributes. We then leverage an offline learned Mahalanobis distance metric on the middle level attributes to yield a final prediction on gender.Finally, We evaluate our method on proposed novel dataset and compare with state-of-art methods based on face specification.

1 Introduction

The machine learning literature provides a versatile perspective on gender recognition problem. Boosted by the face detection and alignment techniques, the past few decades have witnessed the thriving of gender recognition based on face specification [1,2]. It is fairly easy to make a prediction given a detected and aligned face. However, useful face information is often not available in unconstrained environments, due to partial occlusion by hair or sunglasses, low resolution or simply not facing directly to the camera.

To overcome the shortcomings of using facial information alone for gender recognition, we found that clothing appears to be one of the most crucial cues for gender recognition from human's intuition. Surprisingly, gender recognition from clothing is still not well studied yet given the enormous discrepancies of human pose, illumination and background of photos shot in unconstrained environments. In this paper, we propose a novel method that advances the state-of-the-art algorithms for gender recognition on accuracy in case of absent face information. The pipeline of our framework is illustrated in Fig. 2. Given a query image, we segment it into several semantic parts(e.g. blazer, vest, bag) as local regions are more likely than the whole image to be invariant to pose distortion

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 515–523, 2014.

and evite the interference among clothing components. For each region, low-level features are extracted and fed into pre-trained classifiers. By concatenating the posterior probabilities from classifiers, we construct the mid-level attribute which engenders compact visual descriptions.

Our contributions are two-fold:

- We propose an innovative database containing 30,276 images for the masculine and feminine, and parse the photos into regions based on clothing categories. The database has an enormous variety in pose, illumination and background clutter which is an avenue that can be further exploited for other fashion research.
- We propose a novel gender recognition method using mid-level attribute based on semantic clothing parsing. It significantly outperforms state-of-the-art methods in recognition accuracy.

2 Related Work

Gender Recognition. Despite of a rich literature in gender recognition using various features extracted from face [1,2] , a thorough discussion on that topic is beyond the scope of this paper.

In absence of facial information, many efforts have been devoted to gender recognition via other means. [3,4] believe that human bodies can infer gender. Nonetheless, they did not fully utilize the cues in images, and only leveraged features that only depict the shape of the body, like HOG [5]. In our method, we fully utilize cues in image from the perspective of color, texture and shape.

Work that is most similar to ours may be [6], which is significantly competitive in realm of attribute learning on clothing [6,7]. They attempt to generate a list of attributes to describe a clothing image taken in an unconstrained environment, and gender is one of many middle attributes they obtain from low-level features extracted from patches sampled from human poses. However, they only sample patches from upper torso, whereas we found that clothing items from lower body also provide distinctive cues for gender recognition, such as pants and shoes.

Patch-Based Recognition and Retrieval. Dividing the images into several patches, where each part is represented and matched, is a popular technique for object recognition [8,9], as it alleviates the human pose discrepancy. Nevertheless, it contains disturbance owing to occlusion and interaction between clothes regions.

Shrivastava et al.[8] propose a simple but efficient approach that estimates the relative importance of each patch leveraging the notion of 'data-driven uniqueness'. However, this generic approach does not rely on specific visual domain mainly because it divides the image into regular grids without considering the semantic meaning. In our method, we utilize the clothing regions which are parsed by categories and estimate their expected utilities accordingly. For instance, skirts provide a strong prior of femininity, weight of which is significantly larger than that of jeans.

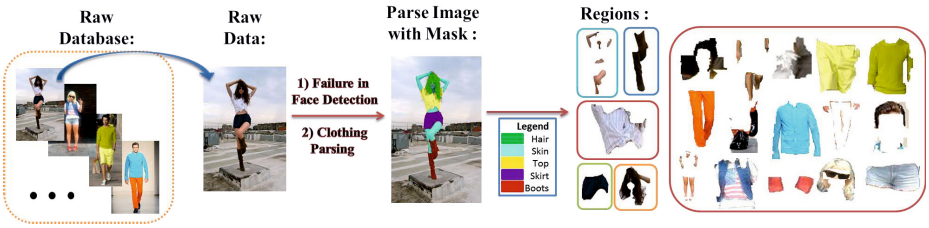

Fig. 1. Pre-process for our database. We first yield face cropping if detection [2] succeeds, otherwise, we employ parsing [10] to get the clothes regions.

3 Database Preparation

[10] proposes a fashion database collected from online fashion sharing websties, but it is mostly restricted to feminine garments, as female are more inclined to post their fashionable grooming on the websites. To balance the ratio, we download masculine images from not only fashion website: Chictopia [11], OSHa'Re [12], but also Google Image Search. Photos in our database are shot in realistic scenes with considerable background clutter and uploaded by fashion lovers who incline to show versatility of garments, accessories and shoes etc. As a result, we collect 15,237 images for women and 15,039 for men for studying gender recognition.

In the offline pre-processing stage, we crop out the face region if the face detection [2] succeeds so as to eliminate the interference, as we are investigating the relation between clothing and gender. We adopt clothing parsing [10] and construct the features from each extracted and cropped region respectively. Yamaguchi et al. [10] address the garment parsing problem by finding similar styles from database which contains tagged fashion images. We worte a script to use their online parsing tool [13] to parse the query image into 56 categories including background, vest, boots and so forth. The parsed results specify which clothing categories they belong to, thus, no manual labeling is required in this stage. Note that we employ the parsed results excluding the background, which contains great discrepancy and interference in our method. Fig. 1 demonstrates the pre-processing of database.

4 Gender Recognition via Attribute Learning

In this section, we outline the basic components of our framework, that is, low-level feature representation and mid-level attribute learning.

Considering the recognition procedure, in [14], it concatenates all features to construct a large one. However, according to both our experiment, the descriptor acquired by concatenating all types of features offers negligible improvements over SVM [15] than that uses color alone. As is illustrated in the work [6], the failure is mainly due to the over-weighted features with high dimension, and engenders similar result to the one with high dimensional feature alone, not to mention the overfit and low efficiency that the high dimensional feature vectors

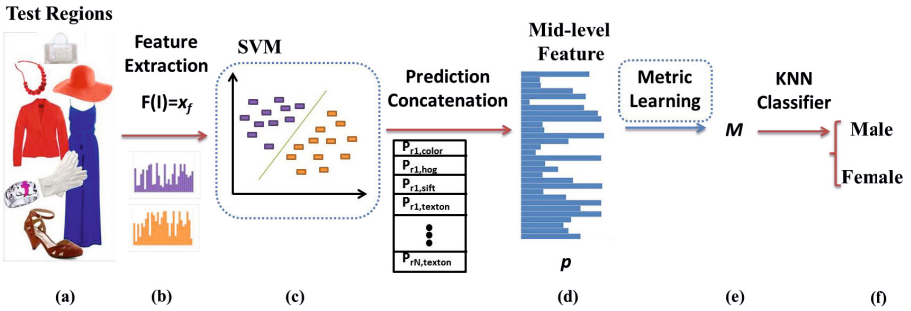

Fig. 2. Pipeline of our proposed approach.(a) Regions parsed in accordance to clothing categories. (b) Diverse features extracted. (c) Features are fed into the classifiers trained offline(bounded by blue dashed box). (d) Results are concatenated as mid-level attribute. (e) Offline metric learning(bounded by blue dashed box). (f) Final prediction is gained based on KNN Classifier

bring about. Thus in our method, we trained mid-level attribute and adopt attribute classifier which will be elaborated in Section 4.3.

4.1 Low-Level Feature Representation

Similar to human's intuition, we represent each clothing region by color, texture ans shape. The first step is to extract low-level features from each parsed semantic regions. For the input image I, we extract the f features $\boldsymbol{x_f}$, where $f \in \mathcal{F} = \{\text{Color, SIFT, Texton, HOG}\}$. For clarity, we describe the representation for each region.

Appearance. Color stands as a critical dimension in encoding appearance. We extract the feature in CIELab space. With the number of bins $l = 10$, $a = 25$, $b = 25$, the normalized values is aggregated by collapsing them into a 6250-dimensional histogram.

SIFT with Bow(Bag-of-words) [16] are computed at salient points, and the descriptors are quantized into $V(= 1024)$ visual words using K-means Clustering.

Texture. Texture discrimination is implemented based on texton [17] with the Gabor filter, which possesses power spectrum of the micro-patterns corresponding to various textures and is proved to be particularly appropriate for the texture representation.

Shape. Local shape is represented by a histogram of edge orientations gradient(HOG) within an image subregion quantized into $K(= 9)$ bins. Each bin in the histogram represents the number of edges that have orientations within a certain angular range.

4.2 Low-Level Feature Classification

With respect to classifier, we adopt linear SVM as it appears to give the best overall performance [4]. We train four linear SVM regressors using LIBSVM [19] with $C = 1.0$ for color, HOG, SIFT and texton and obtain the posterior probabilities instead of binary classification respectively. The classifiers combine the evidence and preferentially use the more informative dimension of feature by assigning more weights to them to determine the classification result. Note that in this training stage, we randomly pick 40% images from the dataset for classification, then we select 35% from the remained data for the metric learning in 4.3.

Each class-label $c \in \mathcal{C} = \{0, 1\}$ represents male or female respectively in the database we collect. Suppose that $\mathcal{R}$ is the set of all regions parsed, $\mathcal{F}$ the set of all types of features. With $Y(I)$ representing the class-label c for image I, given a new testing image I_T, the posterior probability is $P(Y(I_T) = c \mid \boldsymbol{x_f}, r)$, where $r \in \mathcal{R}$ and $f \in \mathcal{F}$.

Concatenating the probabilities $P(Y(I_T) = c \mid \boldsymbol{x_f}, r)$ for all $r \in \mathcal{R}$ and $f \in \mathcal{F}$, we obtain the mid-level attribute, which is further utilized in Section 4.3.

4.3 Metric Learning for Mid-level Attributes and Recognition

In this step, instead of merely accumulate the average of the posterior probabilities as the ultimate result, we assess the expected utility for various regions and, similarly, for different types of features, enabling the various visual features (e.g. color, texture and shape) and cues(e.g. dress, shoes, blazer) to be combined.

The final gender prediction arises from the collection of posterior probability score of each feature classifier and region. We regard the probability scores as elements of mid-level attributes and transform the final prediction to a recognition problem on the new feature vector formed by them, denoted as $\boldsymbol{p}$ which is visualized in Fig. 3. The benefits are two-fold: mid-level attributes could capture local feature distribution well for each semantic region and recognition on mid-level attributes is more robust to noise. Since we extract four types of features from each region(in this paper, the corresponding number of regions is 55), $\boldsymbol{p}$ is a 220-dimension vector.

Since the regions we obtained from clothes parsing are semantic ones, such as shirts, boots, jeans etc, they naturally encode priors on how discriminative they are for gender recognition. We formulate this observation as a metric learning problem on the mid-level attributes vector $\boldsymbol{p}$, where we learn a diagonal matrix $\boldsymbol{W}$ such that $\boldsymbol{Wp}$ correctly reflects the expected utility of each probability score.

We apply a data-driven approach to learn the weight matrix $\boldsymbol{W}$ from all training data. We adopted a metric learning algorithm from [18], where learning $\boldsymbol{W}$ is equivalent to learn a Mahalanobis metric $\boldsymbol{M}$ as proven in [18].

Thus, the metric learning problem is to learn the semidefinite matrix $\boldsymbol{M}$. Given an image I and corresponding ground truth gender $G(I)$ from the training data, we first collect the posterior probability scores and construct $\boldsymbol{p(p_I \mid G(I))}$.

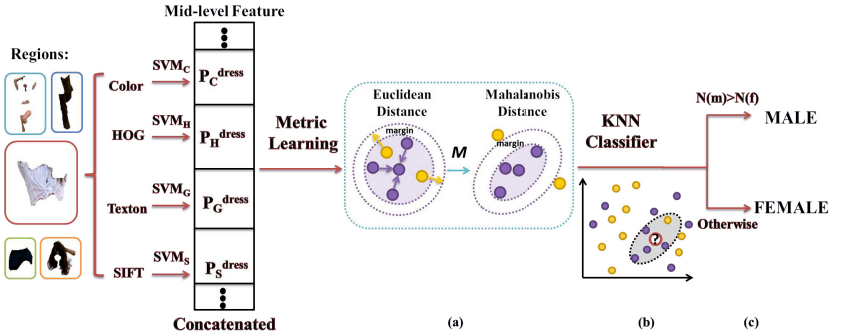

Fig. 3. Construction of mid-level attributes and metric learning. (a)Utilizing metric learning, the weight vector correspondent to mid-level attribute is gained. (b)Employ KNN classifier as the attribute classifier. (c) Final result is obtained.

According to [18], we minimize the following semidefinite programming (SDP):

$$\begin{aligned} \min_{\boldsymbol{M},\xi} \quad & \sum\nolimits_{(\boldsymbol{p_i},\boldsymbol{p_j})\in\mathcal{S}} d_{\boldsymbol{M}}(\boldsymbol{p_i},\boldsymbol{p_j}) + \lambda \sum\nolimits_{(\boldsymbol{p_i},\boldsymbol{p_j},\boldsymbol{p_k})\in\mathcal{R}} \xi_{ijk} \\ \text{subject to} \quad & d_{\boldsymbol{M}}(\boldsymbol{p_i},\boldsymbol{p_k}) - d_{\boldsymbol{M}}(\boldsymbol{p_i},\boldsymbol{p_j}) \geq 1 - \xi_{ijk} \forall (\boldsymbol{p_i},\boldsymbol{p_j},\boldsymbol{p_k}) \in \mathcal{R} \\ & \boldsymbol{M} \succ 0, \xi_{ijk} \geq 0. \end{aligned} \tag{1}$$

where a constraint $(\boldsymbol{p_i},\boldsymbol{p_j},\boldsymbol{p_k}) \in \mathcal{R}$ has the property that $\boldsymbol{p_i}$ and $\boldsymbol{p_j}$ are neighbors from the same gender class and $\boldsymbol{p_i}$ and $\boldsymbol{p_k}$ are of different gender class. We refer interested readers to [18] for further reading.

With the new distance metric at hand, we yield a K-Nearest-Neighbor classifier(empirically set K=101) on all $\boldsymbol{Wp}$ from training data as the final classifier.

5 Experiments

For testing across diversity of pose, illumination and background, we propose a well-suited database and elaborate the collection in Section 3, allowing us to verify the breaking-out results regardless of the variabilities. Our gender recognition method will be compared with the state-of-the-art approaches including Face++ [2], grid-based image matching method [8], pose-adaptive gender prediction [6], and human perception. We study the performance of our framework quantitatively and qualitatively.

Table 1. Performance of gender recognition approaches

	Face Matching [8]	Grid-based Alone [2]	Clothes Attributes [6]	**Ours with Clothes Regions**	Both Face and Clothes Regions	Perception without Face	Perception with Face
Male **False** Positive	28.91%	23.43%	12.35%	**8.33%**	0.40%	5.26%	0.78%
Female **False** Positive	31.71%	8.94%	3.59%	**1.28%**	1.99%	0.80%	0.40%
Total Accuracy	69.72%	83.67%	84.13%	**96.04%**	97.73%	98.54%	99.34%

5.1 Quantitative Results

We utilize our own database, and randomly select 75 % images as the training data, and the remaining ones testing. Table 1 demonstrates the performances of these approaches.

Comparison with Face++. We exploit face++ [2], a mature commercial software, to conduct gender recognition based on face specification. Our method using clothing alone achieves a 20.58% drop in the male false positive rate, 3.59% in the female one compared with face++, and a 26.32% increase of total accuracy. While detailed facial traits provide strong clues about one's identity, face alone is not sufficient for successful gender recognition, especially in unconstrained set photos. Our method with clothing information alone, surpasses face++ significantly.

Comparison with Grid-Based Weighted Inputs. Weights of each grid are trained for grids same as the one in [8]. We adopt $25(=5 \times 5)$ grids, as we notice that when the number of grids reaches 25, the prediction accuracy goes smoothly and steadily. From Table 1, our method achieves a 12.37% increase in accuracy, since simple grid-based patches lack the semantic encoding of the importance of each patch for gender recognition problem as used in our approach.

Comparison with Pose-Adaptive Gender Prediction. We measure the prediction accuracy by using semantic attributes of garments [6]. The recognition result is improved compared with face++. However, without addressing the interaction among clothes items, it is inferior to our proposed method. We utilize our approach with clothing alone(recorded in bold font), and to steer a higher accuracy, with both face and garment information afterwards(shown in column six). Both our method and [6] validate that gender recognition result could be improved by combining clothing information with traditional face-based algorithms.

Comparison with Human Perception. Human perception experiment is conducted as well, with 15 volunteers possessing common knowledge of on fashion and gender, telling the genders in accordance to the images they perceive. Note that test cases shown to the volunteers possess 2 groups: 1) photos with face cropped out, 2) the original photos. The results of human perception is exhibited in the last two columns. With results of human perception standing out, our proposed method is comparable with it.

5.2 Qualitative Results

Fig. 4 displays the failed samples using face++, which are successfully distinguished via our approach. We conclude that, in the real photo shoot, face detection fares poorly in respect of partial occlusion and side face. Besides, wearing sunglasses interferes both detecting and predicting stage. Faces with ambiguous features engender confusions as well.

Fig. 4. Failure cases(which are correctly recognized with clothing information alone) using face specification. The first row demonstrates the failed cases of men, whereas the second women. Left column shows the samples that fail in the very beginning-face detection, and the samples on the right fails in recognition. Failure is mainly on account of partial occlusion by hair, side face, wearing sunglasses, possessing the sexually ambiguous features

6 Conclusions

In short, we propose a robust image representation for gender recognition based on regions parsed by clothing categories. Mutual features are extracted and complemented with weights which are gained in the pre-processing stage. With the discriminative information and metric learning, our method yields superior performance compared to the state-of-the-art algorithms, proving that garments form a profound springboard to attain gender recognition accuracy.

References

1. Makinen, E., Raisamo, R.: Evaluation of gender classification methods with automatically detected and aligned faces. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(3), 541–547 (2008)
2. Incc. M, Face++ research toolkit, `http://www.faceplusplus.com`
3. Cao, L., Dikmen, M., Fu, Y., Huang, T.S.: Gender recognition from body. In: Proceedings of the 16th ACM International Conference on Multimedia, pp. 725–728. ACM (2008)

4. Collins, M., Zhang, J., Miller, P., Wang, H.: Full body image feature representations for gender profiling. In: 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1235–1242. IEEE (2009)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (June 2005)
6. Chen, H., Gallagher, A., Girod, B.: Describing clothing by semantic attributes. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 609–623. Springer, Heidelberg (2012)
7. Bourdev, L., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1543–1550. IEEE (2011)
8. Shrivastava, A., Malisiewicz, T., Gupta, A., Efros, A.A.: Data-driven visual similarity for cross-domain image matching. ACM Transactions on Graphics (TOG) 30(6), 154 (2011)
9. Ferencz, A., Learned-Miller, E.G., Malik, J.: Learning to locate informative features for visual identification. International Journal of Computer Vision 77(1-3), 3–24 (2008)
10. Yamaguchi, K., Kiapour, M.H., Berg, T.L.: Paper Doll Parsing: Retrieving Similar Styles to Parse Clothing Items. In: International Conference on Computer Vision, ICCV (2013)
11. `http://chictopia.com`
12. `http://oshare.com.tw`
13. `http://clothingparsing.com`
14. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 365–372. IEEE (2009)
15. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
16. Uijlings, J.R., Smeulders, A.W., Scha, R.J.: Real-time bag of words, approximately. In: Proceedings of the ACM International Conference on Image and Video Retrieval, p. 6. ACM (2009)
17. Tsai, D.M., Wu, S.K., Chen, M.C.: Optimal Gabor filter design for texture segmentation using stochastic optimization. Image and Vision Computing 19(5), 299–316 (2001)
18. Weinberger, K., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbor classification. Advances in Neural Information Processing Systems 18, 1473 (2006)
19. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), 27 (2011)