Face Gender Classification on Consumer Images in a Multiethnic Environment

Wei Gao and Haizhou Ai

Computer Science and Technology Department,
Tsinghua University, Beijing 100084, China
ahz@mail.tsinghua.edu.cn

Abstract. In this paper, we target at face gender classification on consumer images in a multiethnic environment. The consumer images are much more challenging, since the faces captured in the real situation vary in pose, illumination and expression in a much larger extent than that captured in the constrained environments such as the case of snapshot images. To overcome the non-uniformity, a robust Active Shape Model (ASM) is used for face texture normalization. The probabilistic boosting tree approach is presented which achieves a more accurate classification boundary on consumer images. Besides that, we also take into consideration the ethnic factor in gender classification and prove that ethnicity specific gender classifiers could remarkably improve the gender classification accuracy in a multiethnic environment. Experiments show that our methods achieve better accuracy and robustness on consumer images in a multiethnic environment.

Keywords: Boosting tree, gender classification, multiethnic environment.

1 Introduction

Face vision research has achieved significant advancement in the past decade especially on face detection and face alignment or facial feature location technologies that can readily provide effective tools to extract faces from raw images. With faces having being extracted, demography classification that includes gender, ethnicity and age become an interesting topic due to its potential applications in photo album management, shopping statistics for marketing, visual surveillance, etc. Unlike ethnicity and age estimation, gender classification has attracted more attention in face classification literature since it is the most basic information from a face which human can have a very clear division in perception.

In the early days, most of researches in gender classification are about human's perceiving for gender from a psychology point of view, where the computer is used just as an assistant tool and no automatic gender classification system is developed. More recently neural network methods were firstly used in gender classification. Golomb et al. [1] trained a gender classifier "SexNet" with a two-layer neural network on 90 facial images and achieved an accuracy of 91.9%. Gutta et al. [2] trained a neural network on 3000 faces from FERET dataset and decreased the error rate to 4%.

M. Tistarelli and M.S. Nixon (Eds.): ICB 2009, LNCS 5558, pp. 169–178, 2009.

Balci [3] used eigenfaces and trained a Muti-Layer Perceptron on FERET Dataset to analyze which eigenface contributed to gender classification. Later, Moghaddam et al. [4] used SVM and achieved an accuracy of 96.6% on FERET's 1755 faces which was the best result on this set. However for human, only about 96% in accuracy can be achieved using face information only on this problem according to [5]. BenAbdlkader et al. [6] extracted both local and holistic features and used LDA and SVM on a dataset of about 13,000 faces that achieved 94.2% correct rate. Yang et al. [7] used LDA, SVM and Real AdaBoost respectively on 11,500 snapshots that achieved about 96.5% correct rate. Another interesting work, Lapedriza et al. [9] analyzed the external face's contribution to gender classification and developed a classifier [10] based on both internal and external faces that resulted in 91.7% correct rate on FRGC dataset.

In overview of most of previous works on gender classification, one thing in common is that all those face images used in the experiments are caught in constraint environments, and further they are using 5-fold CV verification method to evaluate performance that implies their test sets have the same distributions as their training sets. So the generalization ability on independent sets is still a problem. Shakhnarovich et al. [8] reported that gender classification on images from internet by AdaBoost algorithm can only achieve 78% accuracy and by SVM can only achieve 77% accuracy. Besides, the ethnic factor is less considered before. The gender classifier trained can not guarantee good generalization ability in a multiethnic environment.

In this paper we target at face gender classification on consumer images of which faces vary greatly in pose, illumination and expression. Active Shape Model (ASM) is used for normalization and a boosting tree is trained on about 10,100 face images from unconstrained environment. Comparative experiments with other methods including SVM and Real AdaBoost on independent test sets are reported to show its effectiveness. To handle a multiethnic environment, we treat face's ethnicity as a latent variable and the ethnicity specific gender classifiers are trained.

The rest of this paper is organized as follows: Section 2 gives an overview of our gender classification system, Section 3 describes the boosting tree classification method, Section 4 gives the gender classifier structure in multiethnic environment, and finally, Section 5 is the conclusion.

2 Gender Classification on Consumer Images

By consumer images we mean those digital photo images caught by popular users of digital cameras. Compared with faces caught from constraint environment, such as snapshots, faces in consumer images are more diverse in resolution, makeup, as well as in illumination, pose and expression (as shown in Fig.1), therefore they are more challenging to deal with in classification. In this situation, preprocessing and normalization become a critical issue.

As for gender classification methods, AdaBoost has been proved very effective in both accuracy and speed in the literature. Since AdaBoost is much faster than SVM, for potential practical applications we choose it to develop a boosting based method for gender classification. In fact, AdaBoost can mine discriminative features automatically from a large set by giving miss-classified samples more attention. Yang et al. [7] and Lapedriza et al. [9][10] showed that boosting algorithm achieved comparative accuracy with SVM in gender classification problem. But the main drawback of this algorithm is

overfitting after over a certain number of iterations which means poor generalization ability in other dataset, especially on those with high intra-class variations. Since faces in consumer images are with great intra-class variations, it is found very difficult to learn a single boosting classifier as in [7] [9] [10] in our experiments, therefore divide and conquer strategy becomes necessary for better performance.

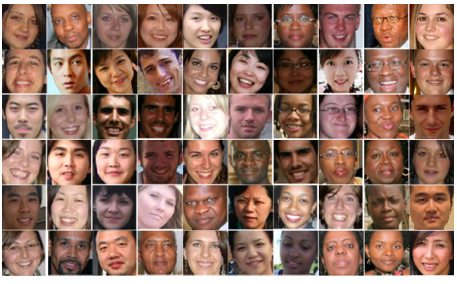

Fig. 1. Faces from Consumer Images

For a flowchart of our gender classification system, see Fig.2 First a face detection algorithm [12] is used to detect faces from consumer images and then a variation of ASM method [13] is used to locate 88 facial feature points for each detected face. For normalization, a shape free face texture is acquired by triangular warping from a shape aligned face to the 32×32 mean face shape. Compared with the conventional eye-center normalization in face recognition approaches, this method eliminates some pose and expression variations.

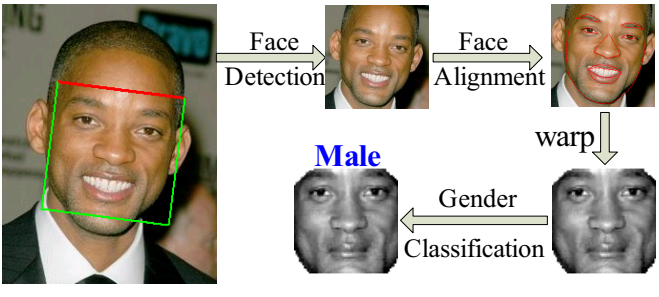

Fig. 2. Face Gender Classification Flowchart

3 Gender Classification by Boosting Tree

3.1 Probabilistic Boosting Tree

The probabilistic boosting tree (PBT) method is originally proposed by Tu [11] to deal with the problem of object categorization in natural scenes. It is a new divide-and-conquer strategy with soft probability boundary. The boosting tree method find the classification boundary step wisely by putting the ambiguous samples to both left

and right sub-trees as shown in Fig.3 (left). Gradually, more similar samples will be sent to sub-tree nodes which results in a reduction of intra-variation. The boosting tree can approach the target posterior distribution by tree expansion.

In the boosting tree structure, each node is a strong classifier trained by AdaBoost algorithm. We adopt the LUT-based Real AdaBoost method in [14] and use simple Haar-like features [16] to construct weak classifiers. After T iterations of learning, the strong classifier has the form:

$$H(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x})$$

where $h_t(\boldsymbol{x})$ is the t-th weak classifier, α_t is the normalize coefficient and $H(\boldsymbol{x})$ is the output confidence.

To construct a boosting tree, the confidence output by the root node is further mapped to probability by sigmoid function as proposed in [15]:

$$q(+1|x) = \frac{\exp\{2H(x)\}}{1+\exp\{2H(x)\}}, \quad q(-1|x) = \frac{\exp\{-2H(x)\}}{1+\exp\{-2H(x)\}}$$

where $q(+1|x), q(-1|x)$ donates the sample's probability to be positive or to be negative respectively. Based on the probability above, we split the training set into sub-trees. This is done by choosing a threshold parameter ε to divided probability into three intervals as shown in Fig.3 (left), that is, the left tree with samples in $[0, \frac{1}{2}-\varepsilon)$, the right tree with samples in $(\frac{1}{2}+\varepsilon, 1]$ and the ambiguous samples in $[\frac{1}{2}-\varepsilon, \frac{1}{2}+\varepsilon]$ will be added into both the left and the right sub-tree (as show in Fig.3 (right)). In practice, instead of using a fixed threshold for every tree nodes as in [11], we choose a variable threshold for each node according to the distribution of samples to make the tree trained more balanced.

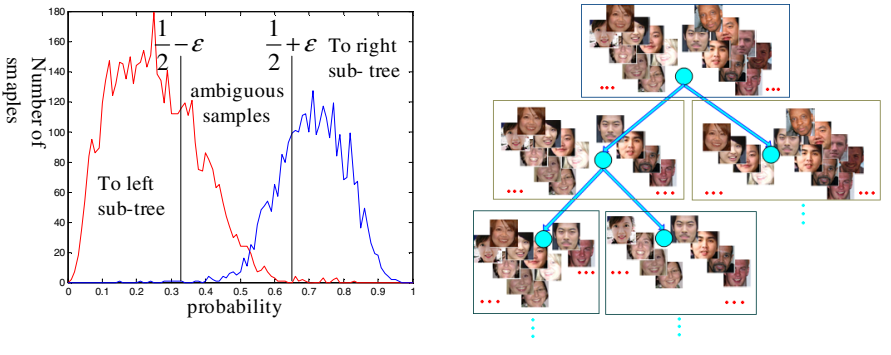

Fig. 3. (left) Histogram of probability distribution of positive and negative samples and three intervals divided. (right) Probabilistic boosting tree structure. Faces in the left and right of the nodes correspond to positive and negative samples.

The above procedure is repeated to construct a tree of which each node is a confidence-rated strong classifier learned by Real AdaBoost algorithm.

With the PBT trained, given a face sample, its normalized face is fed into the root node of the tree to start the decision procedure iteratively. At each node the probability to be a positive sample and that to be a negative sample are denoted as $q_P(+1|x)$ and $q_N(-1|x)$ respectively. And then it will be fed into both its left and right sub-tree to compute its corresponding probabilities $p_{right}(y)$ and $p_{left}(y)$. The final decision is computed as:

$$p(y|x) = q(+1|x)p_{right}(y) + q(-1|x)p_{left}(y).$$

3.2 Experiment Result

Experiments are carried out on two kinds of face image datasets: snapshot datasets, and consumer image datasets. And for each kind of face images, two face sets are established: one for training and 5-fold CV test, while the other is totally independent from the training set which is used to judge the algorithms' generalization ability.

The snapshot training dataset (Snapshots) consists of 15,300 faces from controlled environment. The snapshot faces are all frontal with similar lighting condition and expressions. And the independent snapshot set consists of 1800 faces (Independent Snapshots). The consumer image training dataset consists of about 10,100 Mongoloid faces in real environment (Consumer Images) with significant changes in poses, illumination and expressions. Similarly another independent consumer image dataset is collected which consists of 1,300 faces (Consumer Images). All the face datasets collected above contain nearly equal number of samples for each gender. The boosting tree method is compared with SVM and Real AdaBoost on those datasets.

Table 1 gives both the results on the Snapshot dataset under the 5-fold CV verification protocol and the results tested on two other datasets of which 'All consumer images' means the sum of the two consumer image datasets. The SVM method uses Gaussian kernel. The Real AdaBoost method uses a strong classifier learned after 500 rounds. The boosting tree is composed of 15 strong classifier nodes and its depth is 4. The generalization ability is evaluated on the independent snapshots and consumer image dataset.

Table 1. Results on snapshot dataset under 5-fold CV and results tested on two independent datasets

	5-fold CV	Independent Snapshots	All consumer images
SVM	96.38	87.89	66.37
AdaBoost	96.50	90.41	80.50
PBT	97.13	93.48	82.07

Table 2. Results on Consumer Images under 5-fold CV and results tested on two independent datasets

	5-fold CV	Independent Consumer Images	All snapshots
SVM	90.24	88.13	90.72
AdaBoost	94.12	88.61	92.89
PBT	95.51	92.84	93.71

Table 2 gives both the results on the Consumer Image dataset and the results tested on two other datasets of which 'All snapshots' means the sum of the two snapshot image datasets. As before, the Real AdaBoost method uses a strong classifier learned after 500 rounds, and the PBT with 15 nodes and a depth of 4.

From the above results, we can see from Table 1 that all the three methods achieved comparative performance in snapshot datasets while their generalization ability on the consumer images is bad. However the PBT achieved better generalization ability than the other two methods on independent snapshot dataset. From the Table 2, we can see on the consumer images, PBT's generalization ability remarkably outperforms SVM and Real AdaBoost, and their generalization ability on the snapshot dataset is comparative with the classifier directly trained on snapshots. So, although there are variations between indoor controlled environments and unconstraint environments, the classifier trained on real consumer images from unconstraint environments can achieve better generalization ability. We can conclude that the PBT method can describe the classification boundary more accurately than the other two.

4 Gender Classification in a Multiethnic Environment

Compared with gender classification, ethnicity classification attracts less attention in demography classification. Intuitively ethnicity classification could be done almost in the same way as gender classification technically. But different from gender classification, ethnicity classification is much harder and sometimes even human can not have a very clear division for ethnicity in perception. In literature, G. Shakhnarovich et al. [8] divided ethnicity into two categories: Asian and Non-Asian, while in [7] [18] [19] three categories with Mongoloid, Caucasoid and African were adopted, and in [17] four ethnic labels with Caucasian, South Asian, East Asian, and African are used. In this paper, we use three ethnic labels with Mongoloid, Caucasian and African.

4.1 Generic Gender Classifier

We collect 2400 Mongoloid males and 2500 Mongoloid females, 2400 Caucasoid males and 2400 Caucasoid females, and 1800 African males and 1600 African females from consumer images for training. Another independent test set is collected which contains 400 faces for each ethnicity with half for each gender. We train two kinds of gender classifiers: first we train gender classifier for each ethnicity respectively and the results on test set are show in Table 3; and second, we train a gender classifier using all the males and females in the training set and the results on test set are shown in Table 4. All the gender classifiers in this section are trained in the same way as in Section 3.

Table 3. Gender classifier for each ethnicity respectively (MC, CC and AC mean gender classifier on Mongoloid, Caucasoid and African respectively)

	Mongoloid		Caucasoid		African	
	Male	Female	Male	Female	Male	Female
MC	90.1%	92.9%	95.7%	61.3%	93.5%	50%
CC	78.5%	89.1%	96.8%	88.7%	94%	69%
AC	52%	95.5%	55.5%	96%	94%	82%

Table 4. Generic gender classifier for all ethnicities

Mongoloid		Caucasoid		African	
Male	Female	Male	Female	Male	Female
84.5%	93.5%	86%	93%	95.5%	77%

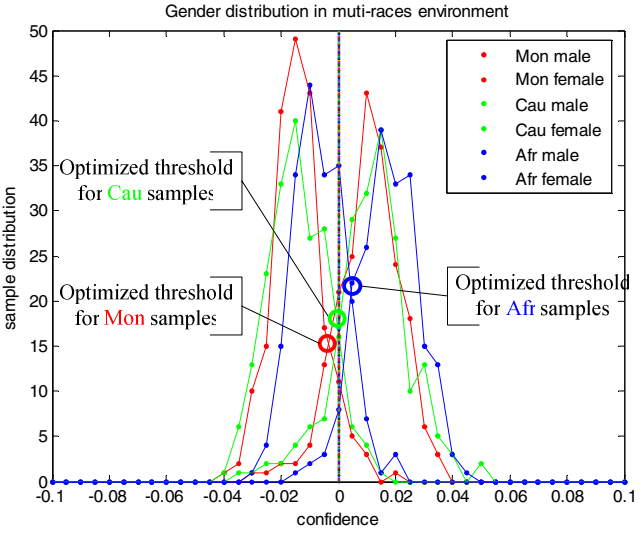

Fig. 4. Confidence distribution for different ethnicity in generic gender classifier (Positive for male and negative for female)

We can conclude from Table 3 that the gender classifier behaves well on the ethnicity it is trained on while can't achieve good results on other ethnicities. When we train a generic gender classifier for all ethnicities, the result as in Table 4 is not as good as training specific gender classifier as in Table 3. This can be explained by Fig.4, in generic gender classifier, we try to find the same threshold for all ethnicity faces, which in fact is not a best decision boundary for each ethnicity. As show in Fig.4, the decision boundary for Africans is apt to male side while the decision boundary for Mongoloid is apt to female side. That is why the generic gender classifier is inclined to classify Africans as males and Mongoloid as females as shown in Table 4.

4.2 Ethnicity Specific Gender Classifier

Enlightened by the analysis in Section 4.1, we propose an ethnicity specific gender classification framework as shown in Fig.5 for multiethnic environment. In the new framework, the ethnicity is treated as a latent variable for gender classification. We can formalize the procedure as:

$$P(G \mid F) = \sum_{E} P(G \mid E, F) P(E \mid F)$$

where G, E and F represent gender, ethnicity and face respectively.

We trained an ethnicity classifier with samples collected in Section 4.1 using AdaBoost.MH [20] and Haar-like features [16]. Gender classifiers on each ethnicity are from Section 4.1. The results of ethnicity specific gender classifier are compared with the generic gender classifier in Table 5. We can see that the ethnicity specific gender classifier performs better than the generic gender classifier, especially on Mongoloid males and African females, which is consistent with analysis of Fig.4. This experiment hints that faces from different ethnicity have different gender feature and in a multiethnic environment, gender classifier could be better by taking ethnicity as a latent variable. Some results are shown in Fig.6.

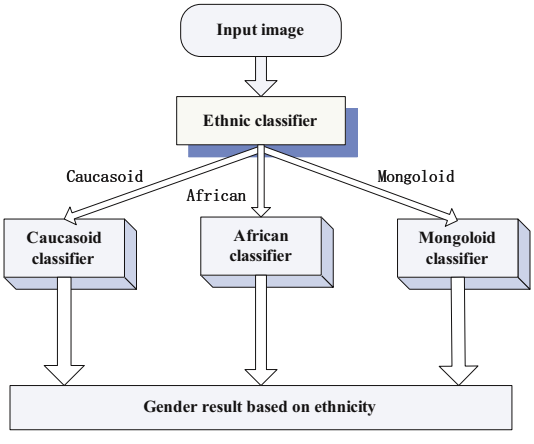

Fig. 5. Ethnicity specific gender classification framework

Table 5. Comparison of Generic Gender Classifier (GGC) and Ethnicity Specific Gender Classifier (ESGC)

	Mongoloid		Caucasoid		African	
	male	female	male	female	male	female
GGC	84.5%	93.5%	86%	93%	95.5%	77%
ESGC	89%	93%	86.3%	96.6%	94%	82%

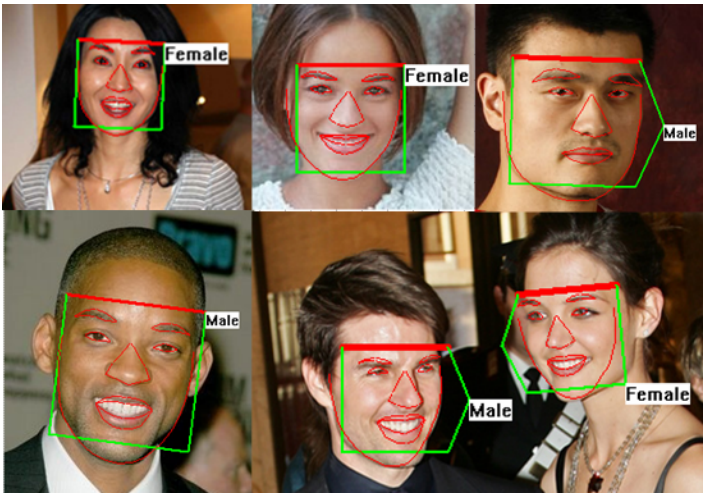

Fig. 6. Gender classification results on consumer images in multiethnic environment

5 Conclusion

In this paper, a PBT approach for face gender classification on consumer images is presented. Faces on consumer images vary greatly in pose, illumination and expression that make it much more difficult than in the constrained environments. In this approach, Active Shape Model (ASM) is used for normalization and a PBT is trained for classification by which through divide and conquer strategy a more accurate classification boundary on consumer images is achieved. Experiments on both snapshots and consumer images show that the PBT method is better than the SVM and Real AdaBoost methods.

We also discussed the ethnicity factor in gender classification experimentally, to our best knowledge there is no such work before. We find that faces from different ethnicity have different gender feature, and gender classifier trained on a specific ethnicity could not get good generalization ability on other ethnicities. Finally, we improve the performance of gender classification in a multiethnic environment by treating ethnicity as a latent variable.

However, currently we can only deal with frontal or near frontal faces from consumer images. And the accuracy of gender classifier on Africans is not as high as on Mongoloid and Caucasoid. Another issue we have not considered is the impact of age on the face gender classification. Those are our future work.

Acknowledgement

This work is supported by National Science Foundation of China under grant No.60673107, and it is also supported by a grant from HP Corporation.

References

1. Golomb, B.A., Lawrence, D.T., Sejnowski, T.J.: SEXNET: A Neural Network Identifies Sex from Human Faces. In: NIPS 1990 (1990)
2. Gutta, S., Wechsler, H., Phillips, P.J.: Gender and Ethnic Classification of Face Images. In: FG 1998 (1998)
3. Balci, K., Atalay, V.: PCA for Gender Estimation: Which Eigenvectors Contribute? In: ICPR 2002 (2002)
4. Moghaddam, B., Yang, M.H.: Learning Gender with Support Faces. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(5) (May 2002)
5. Hayashi, J., Yasumoto, M., Ito, H., Koshimizu, H.: Age and Gender Estimation based on Wrinkle Texture. In: ICPR 2002 (2002)
6. BenAbdelkader, C., Griffin, P.: A Local Region-based Approach to Gender Classification from Face Images. In: CVPR 2005 (2005)
7. Yang, Z., Li, M., Ai, H.: An Experimental Study on Automatic Face Gender Classification. In: ICPR 2006 (2006)
8. Shakhnarovich, G., Viola, P.A., Moghaddam, B.: A Unified Learning Framework for Real Time Face Detection and Classification. In: AFG 2002 (2002)
9. Lapedriza, A., Masip, D., Vitrià J.: Are External Face Features Useful for Automatic Face Classification. In: CVPR 2005 (2005)
10. Lapedriza, A., Manuel, M.J., Jiménez, J.M., Vitrià, J.: Gender Recognition in Non Controlled Environments. In: ICPR 2006 (2006)
11. Tu, Z.: Probabilistic Boosting-Tree: Learning Discriminative Models for Classification, Recognition, and Clustering. In: ICCV 2005 (2005)
12. Huang, C., Ai, H., Wu, B., Lao, S.: Boosting nested cascade detector for multi-view face detection. In: ICPR 2004 (2004)
13. Zhang, L., Ai, H., Xin, S., Huang, C., Tsukiji, S., Lao, S.: Robust Face Alignment Based on Local Texture Classifiers. In: ICIP 2005 (2005)
14. Wu, B., Ai, H., Huang, C.: LUT-Based AdaBoost for Gender Classification. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688. Springer, Heidelberg (2003)
15. Schapire, R.E., Singer, Y.: Improved Boosting Algorithms Using Confidence-rated Predictions. Machine Learning 37, 297–336 (1999)
16. Viola, P., Jones, M.: Fast Multi-view Face Detection. In: Proc. of CVPR (2001)
17. Gutta, S., Huang, J.R., Jonathon, P., Wechsler, H.: Mixture of Experts for Classification of Gender, Ethnic Origin, and Pose of Human Faces. IEEE Transactions on Neural Networks
18. Yang, Z., Ai, H.: Demographic classification with local binary patterns. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 464–473. Springer, Heidelberg (2007)
19. Hosoi, S., Takikawa, E., Kawade, M.: Ethnicity Estimation with Facial Images. In: FG 2004 (2004)
20. Schapire, R.E., Singer, Y.: Improved Boosting Algorithms Using Confidence-rated Predictions. Machine Learning (1999)