On the Generalization Power of Face and Gait Gender Recognition Methods

Yu Guan*, Xingjie Wei, and Chang-Tsun Li,
Department of Computer Science, University of Warwick,
Gibbet Hill Road, Coventry, CV4 7AL, UK
{g.yu, x.wei, c-t.li}@warwick.ac.uk

Abstract

Human face/gait-based gender recognition has been intensively studied by the previous literatures, yet most of them are based on the same database. Although nearly perfect gender recognition rates can be achieved in the same face/gait dataset, they assume a closed-world and neglect the problems caused by dataset bias. Real-world human gender recognition system should be dataset-independent, i.e., it can be trained on one face/gait dataset and tested on another. In this paper, we test several popular face/gait-based gender recognition algorithms in a cross-dataset manner. The recognition rates decrease significantly and some of them are only slightly better than random guess. These observations suggest that the generalization power of conventional algorithms is less satisfied, and highlight the need for further research on face/gait-based gender recognition for real-world applications.

Index Terms- Human gender recognition, database bias, face, gait, generalization power, biometrics

INTRODUCTION

Human gender recognition can be used in a wide range of real-world applications such as video surveillance. In terms of biometric traits, face and gait may be the most important modalities that can be used for gender classification (Ng *et al*., 2012). Although gender recognition are intensively studied by the previous literatures, most of them are based on a single dataset (Baluja & Yang, 2007; Li *et al.*, 2008; Moghaddam *et al.*, 2002; Shan *et al.*, 2008; Wang *et al.*, 2010). Unlike the human identification systems, gender recognition should be able to be performed across different datasets in real-world scenarios (Ng *et al*., 2012). Each dataset has its own database bias due to its own unique data collection environments, yet in the context of face/gait-based gender recognition, most of the previous works simply neglect this issue. Although several popular methods like SVM (Moghaddam *et al.*, 2002; Li *et al.*, 2008) , AdaBoost (Baluja & Yang,2007; Wang *et al.*, 2010), PCA+LDA (Shan *et al.*, 2008; Chang *et al.*, 2009) can yield high performance on the same dataset (referred to as intra-dataset), they are seldom evaluated in a cross-dataset manner. In this work, we test these algorithms to see whether they are robust enough against the bias from different datasets. Fig.1 demonstrates several face images from 5 different face datasets while Fig.2 provides some Gait Energy Images (GEI, i.e., average gait silhouette over a gait cycle (Han & Bhanu,2006)) from 2 different gait datasets, can you tell the bias pattern for each group of face/gait images in Fig.1/Fig.2?

Figure 1: Cropped images from the face datasets: (a) AR, (b) FERET, (c) FRGC, (d) LFW, and (e) TFWM.

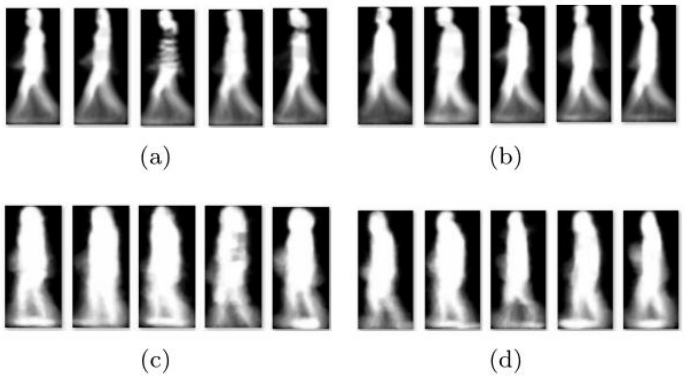


Figure 2: GEI samples: (a) female samples from CASIA-B dataset, (b) male samples from CASIA-B dataset, (c) female samples from USF dataset, (d) male samples from USF dataset.

EXPERIMENTAL SETUP AND RESULTS ANALYSIS

In the previous works, high performance can be achieved when conventional machine learning methods like SVM, AdaBoost, PCA+LDA are used for face/gait-based gender recognition. However, gender is a cue across all datasets and should be independent of specific face/gait dataset. Since each dataset has its own bias (Torralba & Efros, 2011) due to its own unique data collection environments, in this work by testing several popular algorithms in a cross-dataset manner, we aim to evaluate the generalization power of these methods, which are important for practical applications. Correct Classification Rate (CCR) is used to measure the performance.

Gender Recognition by Face

We conduct a series of cross-dataset experiments on four representative face datasets (i.e., AR (Martinez & Benavente, 1998), FERET (Phillips *et al.*,2000), FRGC (Phillips *et al.*, 2005), and LFW

(Huang *et al.*, 2007)) in gender recognition literatures and one dataset (i.e., TheFaceWeMake(TFWM)) collected under realistic conditions by Dexter Miranda (2010). All images are aligned by manually annotated landmarks and cropped to 64×64 pixels. Some of the cropped example images are shown as Fig.1. The detailed information for each dataset is described as follows:

1. AR: contains more than 4,000 frontal view images of 70 males and 56 females with different facial expressions, illumination conditions and occlusions.
2. FERET: contains 14,126 images of 1,199 individuals with different poses, illuminations and expressions.
3. FRGC: contains 44,832 still images of 688 individuals with different illuminations and expressions. Images are taken under controlled and uncontrolled environments.
4. LFW: contains more than 13,000 images of faces collected from the web. Variations include changes in pose, illumination and occlusions.
5. TFWM : contains more than 2,000 frontal view images of strangers on the streets taken under outdoor environment with uncontrolled illumination.

Table 1: CCRs (%) using SVM

Train \ Test	AR	FERET	FRGC	LFW	TFWM
AR	**92.5**	66.7	60.9	59.0	72.4
FERET	77.9	**88.1**	60.0	64.9	69.6
FRGC	61.4	64.2	**78.0**	70.1	65.8
LFW	75.3	70.9	64.7	**76.6**	59.6
TFWM	81.2	56.9	64.4	59.9	**81.7**

Table 2: CCRs (%) using AdaBoost

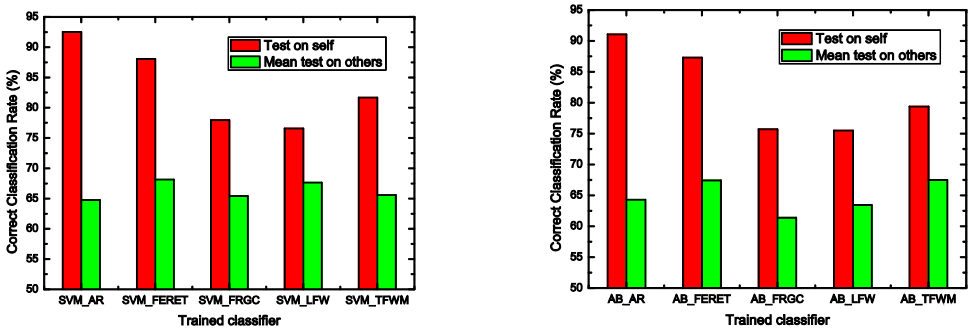
Train \ Test	AR	FERET	FRGC	LFW	TFWM
AR	**91.0**	70.6	61.1	57.8	67.7
FERET	76.9	**87.3**	67.5	61.2	64.0
FRGC	61.1	60.1	**75.7**	61.8	62.7
LFW	60.6	67.4	64.5	**75.5**	61.3
TFWM	77.7	66.0	63.4	63.0	**79.4**

For feature extraction, we use LBP ($LBP_{8,1}^{u2}$ operator), which is one of the most popular descriptors for gender recognition. We choose two classical methods: SVM and AdaBoost for the gender classification tasks. Specifically, we employ the RBF kernel in SVM and 35 weak classifiers in AdaBoost. Both methods are set with the default parameters. We select 1400 images (700 female images and 700 male images) from each dataset. To maintain the gender balance, we randomly choose half of the males and the females as training set while the rest are used as test set. Besides, the same individual does not appear on both training and test sets. Notice that this is the maximum size possible across all datasets (We do not select the occluded images in AR database since there are only two types of occlusions: sunglasses and scarves for all subjects, which may cause the occlusion bias during gender classification). For each experiment, we perform the training process using one dataset and then test it on other datasets. Each experiment is repeated 10 times with the average CCR reported in Table 1 and Table 2 for these two methods.

It is worth noting that our goal is not to beat the best rate for gender recognition in the literatures

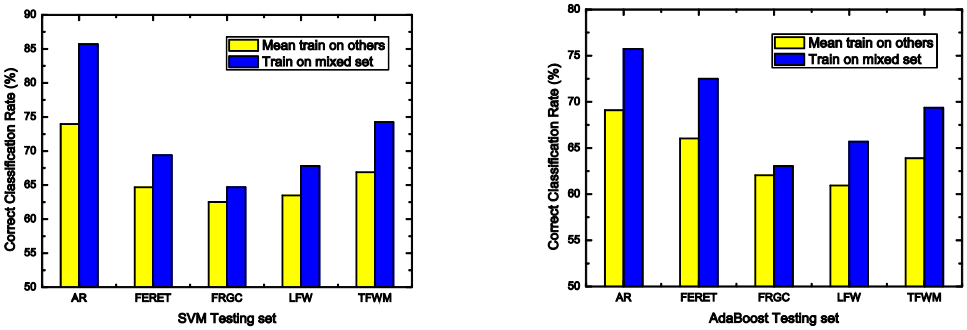
but to evaluate the cross-dataset generalization ability of each trained classifier. So the differences of performance using different training sets are more meaningful than the actual performance rates. It can be found that in each row (of Table 1 and Table 2), the best CCR is achieved when training and testing on the same dataset (i.e., intra-dataset). There is a significant drop when testing on other datasets (i.e., cross-dataset cases) both for SVM and AdaBoost (also see Fig.3). For example, for the classifier trained on AR database, the performance drops nearly 30% for test sets from other datasets. From this we can see, when an algorithm is claimed to achieve satisfied (or, even the best) CCR in one dataset, the cross-dataset testing should be adopted to evaluate the its generalization power against the dataset bias. This is quite important but has not attracted enough attentions in the face-based gender recognition community yet.

In Fig. 3a, we observe that the performance drop is less significant when the classifiers are trained on the LFW dataset. The images from the LFW dataset are captured under uncontrolled conditions with large diversity, which may reduce the dataset bias to some extent. Motivated by this, we conduct five-fold cross validation experiments. One fold (dataset) is used for testing and the others from a mixed dataset of 696 images (174 images from each dataset maintaining the gender balance) for training. We report the recognition results in Fig.4. Compared with the performance based on the training set from a single dataset, the performance is significantly improved when using a mixed-dataset training set. The experimental results suggest that increasing the diversity (e.g., age, ethnicity, illumination conditions, facial expressions, etc.) of training images is a possible way to enhance the performance for face-based gender recognition.



(a) SVM (b) AdaBoost

Figure 3: CCR drops in the cross-dataset cases. SVM_AR denotes that SVM is trained on the AR dataset. The red bars indicate the CCRs corresponding to the intra-dataset experiments while the green bars indicate the average CCRs corresponding to the cross-dataset experiments.



(a) SVM (b) AdaBoost

Figure 4: CCR increases when training on the mixed-dataset. The yellow bars indicate the average CCR of training on single dataset while the blue bars indicate the CCR of training on the mixed-dataset.

Gender Recognition by Gait

We conduct some cross-dataset experiments on two popular gait databases, namely USF dataset (Sarkar *et al.*, 2005) and CASIA-B dataset (Yu *et al.*, 2006). USF is an outdoor dataset with lower segmentation quality while CASIA-B is an indoor dataset with higher segmentation quality. We use GEI as feature template and several GEI samples are illustrated in Fig.2. The GEIs are available in the USF dataset and recently, Zheng *et al.* (2011) made the CASIA-B GEIs available and we only use the ones from the side view (view 90˚) in this work. The GEIs from these two datasets are normalized to the size of 128×60 pixels. In the CASIA-B dataset, for gender balance we choose 31 males and 31 females in the experiments. Since in CASIA-B, most of the subjects are young people with ages between 20 and 30 (Yu *et al.*, 2006), to minimize the influence of age, we also select 31 males and 31 females with the age range between 19 and 30 from the USF dataset.

We test two popular gait-based gender recognition algorithms, namely PCA+LDA and SVM. For comparison, we also employ these two methods on the conventional intra-dataset experiments, based on the Leave One Out Cross Validation (LOOCV) scheme. That is, we use 1 pair of male and female as test set while the rest 30 pairs are used as training set. The process is repeated 31 times for each pair in turn to be tested. Following the experimental setup of some previous works, for PCA+LDA, Nearest Neighbour (NN) rule is used (Shan *et al.*, 2008) while for SVM, we use the linear kernel (Li *et al.*, 2008; Yu *et al.*, 2009). We report the experimental results in Table 3 and Table 4 for both methods, and we can observe that the performance on cross-dataset is significantly lower than the performance on conventional intra-dataset.

In the cross-dataset experiments, there are large number bias factors that can degrade the performance. Although we restrict the age bias by selecting young individuals only, other bias factors such as, clothing, ethnicity, segmentation quality, etc. are not controlled. In (Yu *et al.*, 2009), based on the well-segmented indoor CASIA-B and the well-segmented indoor Soton dataset (Shutler *et al.*, 2002), Yu *et al.* studied the cross-race (Asians and Europeans) gender recognition, and the experimental results suggested that ethnicity and basic clothes types are not the main bias factors that

can affect the gender recognition performance. Given that, maybe the main factor is the segmentation quality. Compared with CASIA-B dataset, subjects in USF dataset are taken from the outdoor environments under the influences of illumination and complicated background, and this would result in imprecise segmentation quality. We can see such effect in Fig.2 from an intuitive perspective. Compared with the well-segmented CASIA-B GEIs (Fig.2a-b), the USF GEIs in Fig.2c-d suffer from the shadows and imprecise segmented body boundaries, which may cause high level of dataset bias. In the cross-dataset cases, we can see the performance become much worse than intra-dataset cases, i.e., only about 10% higher than random guess.

Table 3: CCRs (%) using PCA+LDA

Train \ Test	CASIA-B	USF
CASIA-B	**96.8**	61.7
USF	73.4	**78.2**

Table 3: CCRs (%) using SVM

Train \ Test	CASIA-B	USF
CASIA-B	**96.0**	58.5
USF	62.5	**76.2**

CONCLUSIONS

In this paper, we conduct experiments based on several popular face/gait-based gender recognition algorithms in a cross-dataset manner. The experimental results suggest that the performance can be significantly affected by the dataset bias, since each dataset has its own unique data collection environments. Although the performance of face-based gender recognition can be improved by increasing the diversity of the training set, it is still much worse than the ones based on intra-dataset scenarios. Although several pervious works claimed that they have nearly perfect gender recognition rates, they may not generalize well to data in unknown environments, and thus less practical in real-world applications. Clearly this work is only a start to broader research which will require greater attention in the future in the area of human gender recognition.

REFERENCES

Baluja, S. & Rowley, H. A. (2007). Boosting sex identification performance. *International Journal of Computer Vision*, 71(1), 111-119.

Chang, C.-C. & Lin, C.-J. (2011). LIBSVM: A library for support vector Machines. *ACM Transaction on Intelligent Systems and Technology*, 2(3), 27:1-27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chang, P.-C., Tien, M.-C., Wu, J.-L., & Hu, C.-S. (2009) Real-time gender classification from human gait for arbitrary view angles," *Proceedings of IEEE International Symposium on Multimedia*, 88-95.

Freund, Y. and Schapire, R. E. (1995) A decision-theoretic generalization of on-line learning and an application to boosting," *Proceedings of the Second European Conference on Computational Learning Theory*, 23-37.

Han, J. & Bhanu, B. (2006). Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 28(2), 316-322.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E.(2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments (Tech. Rep. No. 07-49), US, Amherst ,University of Massachusetts.

Li, X., Maybank, S., Yan, S., Tao, D., &Xu, D. (2008). Gait components and their application to gender recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews,* 38(2), 145-155.

Martinez, A. and Benavente, R.(1998) The ar face database (Tech. Rep. No.24), US, Purdue University, Robot Vision Lab.

Miranda, D. The face we make. http://thefacewemake.org/

Moghaddam, B. & Yang, M.-H. (2002). Learning gender with support faces.*IEEE Transactions on Pattern Analysis and Machine Intelligence,* 24(5), 707-711.

Ng, C. B., Tay, Y. H. & Goi, B.-M. (2012). Vision-based Human Gender Recognition: A Survey. *CoRR,* abs/1204.1611.

Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Chang, J., Hoffman, K., Marques, J., Min, J., & Worek, W. (2005). Overview of the face recognition grand challenge," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,*1, 947-954.

Phillips, P., Moon, H., Rizvi, S., & Rauss, P.(2000). The feret evaluation methodology for face recognition algorithms, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22(10), 1090-1104.

Sarkar, S., Phillips, P., Liu, Z., Vega, I., Grother, P., & Bowyer, K.(2005). The humanid gait challenge problem: data sets, performance, and analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 27(2),162-177.

Shan, C., Gong, S., & McOwan, P. W. (2008). Fusing gait and face cues for human gender recognition. Neurocomputing, 71, 1931-1938.

Shutler, J., Grant, M., Nixon, M. S., & Carter, J. N. (2002). On a large sequence-based human gait database. *Proceedings of International Conference on Recent Advances in Soft Computing*, 66-72.

Torralba, A. & Efros, A.(2011). Unbiased look at dataset bias. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,*1521-1528.

Wang, J.-G., Li, J., Lee, C. Y., & Yau, W.-Y.(2010). Dense sift and gabor descriptors-based face representation with applications to gender recognition. *Proceedings of International Conference on Control Automation Robotics Vision*, 1860 -1864.

Yu, S., Tan, T., Huang, K., Jia, K., &Wu, X. (2009). A study on gait-based gender classification," *IEEE Transactions on Image Processing*, 18(8), 1905-1910.

Yu, S., Tan, D., & Tan, T. (2006). A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," *Proceedings of International Conference on Pattern Recognition*, 4, 441-444.

Zheng, S., Zhang, J., Huang, K., He, R., and Tan, T. (2011). Robust view transformation model for gait recognition. *Proceedings of IEEE International Conference on Image Processing*, 2073-2076.