PFW: a face database in the wild for studying face identification and verification in uncontrolled environment

Hai Wang
Department of Computer Science and Engineering
Pohang University of Science and Technology
Pohang, Korea 790784
Email: haiwang@postech.ac.kr

Bongnam Kang
Department of Computer Science and Engineering
Pohang University of Science and Technology
Pohang, Korea 790784
Email: bnkang@postech.ac.kr

Daijin Kim
Department of Computer Science and Engineering
Pohang University of Science and Technology
Pohang, Korea 790784
Email: dkim@postech.ac.kr

***Abstract*—To train and evaluate various face recognition algorithms, quite many databases have been created. But most of them have been created under controlled conditions to study the specific variations of the face recognition problem. These variations include position, pose, lighting, background, camera quality and gender. But in real environment, there are also many applications in which there is little or no control over such variations. Labeled Faces in the Wild, a database has been provided to study the latter, unconstrained face recognition problem. However, LFW is proposed for face verification problem, while we observe that a good verification performance cannot guarantee a good identification performance in real situation. Further, the face images in LFW are not sufficient for training to get a state of the art performance. PFW, POS Faces in the Wild, on the contrast, is a large database which can be served both for evaluating face verification and face identification algorithms. Specifically, PFW contains a certain number of identities and each identity contains quite many images, thus make it suitable both for large scale supervised and semi supervised training. In this paper, we also provide some rules for evaluating the identification algorithm performance in real environment. To the best of our knowledge, our database is the first public available large face data set proposed for face identification in unconstrained environment.**

I. Introduction

First of all, we would like to emphasize that this paper release a large face image database for studying the problem of unconstrained face recognition, both for face verification and identification. The database is currently under the post processing step, after fully finished this work, we will release this data set.

Although face recognition has been a highlight research topic for decades, but the concept of face recognition has some ambiguity. Generally speaking, usually, when we refer to face recognition, it means the following two problems:

1) Given two face pictures, decide whether the two face images represent the same individual or not, usually we call this face verification problem.
2) Given a face picture, find the corresponding identity from a gallery which contains a set of face pictures, this problem is usually referred as the face identification problem. In face identification problem, it contains both open set test and close set test. In close set test, we simply find the most similar face to the querying face, while in open set test, despite finding the nearest identity, we need to decide whether the face image is registered in the gallery or not, usually this requires a performance trade off between the false reject rate and identification rate. Many literatures report the performance in close set test, and not too much research related to the open set problem, while in real face recognition application such security, open set test is commonly widely used.

Our database, which we called POS Labeled Faces in the Wild (PFW)[1], is designed to touch both these two problems, face verification and identification. Generally speaking, we intend to provide an extremely large set contains relatively unconstrained face images. Our face images have a large range of the variation seen in everyday life. This includes variation in pose, lighting, expression, background, race, ethnicity, age, gender, clothing, hairstyles, camera quality, color saturation, focus and other variations. Compared with the well known Labeled Faces in the Wild(LFW) [12], the total face images number in our database is much larger. Further, our database contains more face images per identity which allows our database can be used for face identification problem, and sufficient faces per person make our data set suitable for supervised training which is a supplementary for the LFW database.

Before proceeding with the details of the database, we present some summary statistics and properties of our database:

[1] All the images collected from web, and the copyright belongs to the original

978-1-4799-2190-4/13 $31.00 © 2013 IEEE
DOI 10.1109/ACPR.2013.53

Fig. 1. Some Faces in our database

1) The database contains roughly 85,000 face images, and about 3,200 identities. All identities have more than one image, most of them have more than 20 images.
2) The label for each image is provided, which allows the potential users use the database to train different algorithms, both for supervised learning and semi supervised learning.

Some example face images are shown in the Fig.1:

The specific details are given in the remainder of the paper, which are organized as follows: In Section II, we discuss other databases. In Section III, we demonstrate that face identification and face verification problem is different. Section IV describe the structure of our database and its intended use. Conclusion are made in Section V.

II. Related Databases

There are a large number of face databases available to researchers in face recognition. We list part of them in Table. I. Although these databases range in size, scope and purpose, but most of them were proposed to study face recognition in controlled environment, i.e., to study the influence of one specific parameter to face recognition performance. As analyzed in [12] [14], face recognition in real environment has become more and more important. Under this trend, a database contains a set of images originated from every day life is desired. To the best of our knowledge, until now, there are only two face databases for uncontrolled environment, namely, Caltech 10000 web Faces [3]and LFW, we will analyze them one by one.

Labeled Faces in the Wild [12] This is the widely used database for face verification problem. The database contains images of 5749 different individuals. Among these, 1680 people have two or more images, the remaining 4069 people have just a single image. It has two different views, one is for training to find the best parameter and another one is for reporting the performance. Also, it supports two paradigm, i.e., restricted protocol and unrestricted protocol. LFW is suitable for studying face verification, but as we will show later, some algorithms show good face verification performance on LFW but they show poor identification performance in real environment. Even in case of constrained environment, it has a poorer performance than the baseline obtained from the widely used LBP [21] based algorithms. From this point, LFW is not perfect to study face recognition problem, also, [12] admits that LFW is proposed just for verification problem, thus some databases for face identification are still needed for a fair comparison between different face recognition algorithms. If we look into details in LFW, we can find that, for lots of person in LFW, they just has one face image, and this made LFW impossible to study face identification problem. Further, from a certain number of reports, they suggest that a large external training data set usually guarantee a better face verification performance, a data set contains larger face images is mandatory to achieve state of the art performance [25][26]. Third, the data structure of LFW database limits the training algorithm, lots of person has just one image, this data set is just suitable for semi supervised training algorithm, not suitable for supervised learning algorithm.

TABLE I. Some Databases

DB Name	Identity No.	Image No.
AR DB [1]	126	4000
ORL DB [2]	40	400
CAS PEAL [4]	1040	99594
FRGC DB [5]	466	50000
FERET DB [6]	1199	14126
M2VTS [7]	37	185
NIST DB [8]	1573	3248
CMU-PIE [9]	68	41368
Yale Face DB [10]	15	165
Yale Face DB B [11]	10	5760
Caltech Web Faces [3]	10000	10000
LFW [12]	5006	13250
PF07 [13]	5006	13250
BioID Face DB [18]	23	1521
Georgia Face DB [19]	50	750
Oulu Face DB [20]	125	2000

Caltech 10000 Web Faces [3]. The Caltech 10000 Web Faces database also provides a very broad distribution of faces. The distribution of faces included in the Caltech collection is similar to the distribution of faces in LFW, and the Caltech data set includes significantly background area in the target images. Due to this characteristic, the Caltech database is geared more toward face detection and alignment rather than face recognition. It provides the position of four facial features, but does not give the identity of individuals. Thus, it is not particularly suitable for face recognition experiments.

PF07 Database [13]. The POSTECH Faces database has 200 identities and 64000 face images totally. But all these face images are captured in the controlled environment and the facial variation is not so significant. Further, all the identities are the Asian people. Using this face database to study the specific parameter in face recognition is preferred, however, PF07 database not suitable for study the unconstrained face recognition problem.

III. Face Verification is really enough ?

Since the release of LFW, it has been a widely used benchmark to evaluate the face recognition algorithms, lots of researcher report their performance on LFW database. But to the best of our knowledge, until now, no paper investigate the relationship between verification performance and identification performance. [14] points out that some algorithms achieve state of the art verification performance on LFW database but get poor performance in identification problem in real environment, but they donnot investigate the reason behind this phenomena. During our experiment, we also find this interesting phenomena. Two experimental results are given in Table.II and Table.III.

TABLE II. EVALUATION RESULT ON LFW

Test Scenario	**CSML** [15]	**ULBP** [21]
Verification	85%	74%
Identification	16%	18%

TABLE III. EVALUATION RESULT ON KISA

Test Scenario	**CSML** [15]	**A-SURF** [16]
Verification	95%	92%
Identification	90.8%	96.7%

In Table.II, the verification performance is evaluated on LFW, and the identification performance is tested from a small set of LFW in View 1 (no overlap with the training set)[2]. From Table.II, we can find that the CSML [15] algorithm show poor performance on identification problem in the open set test. In Table.III, both the verification and identification performance are measured on the KISA database (which is a large commercial database captured in controlled environment). From Table.III, we can find Affine-SURF [16] show good performance on face identification problem, but has poorer performance on face verification problem. We also find this phenomena in some other metric learning based face verification approaches [22][23][24], this differs from what we have expected, previously, we think that a state of the art verification algorithm should also show good performance on identification problem. Why this phenomena happen? A brief explanation is that in verification problem, we design the classifier focusing on the binary classification accuracy while neglect the true similarity of two faces. In our experiment, we check the similarity distribution of intra pairs and extra pairs, we find that similarities of the intra pairs center on certain value while similarities of the extra pairs center on another certain value, both of these two distribution have a small variation. If we just consider the binary classification performance, this is the desired result we want to obtain. But in case of identification, since after learning, the similarities are transformed to guarantee the binary classification, the transformed similarity can not represent the ground truth similarity observed by our human, what's more, the similarities become totally disordered. In identification problem, we find the most similar face through ranking the similarity, the disordered similarities might result in some wrong identification performance. Additionally, we find that if the objective function of a learning algorithm focus on optimizing the distance difference between intra pairs and extra pairs regardless of the statistical result of the training data, usually it will result in a poorer identification performance. If the learning algorithm just focus on obtaining some natural statistical result from the training set without considering optimizing the intra and extra pairs similarities, usually the performance decreasing in identification is smaller or no performance drop at all.

IV. INTENDED USE

As introduced in previous section, our data set can be used for two approaches, one is for face verification problem and another one is for face identification problem. First, we emphasize that our data set has no overlapping face images with the LFW data set.

A. Face Verification

In face verification problem, [12] has given a good example for evaluating the performance, since lots of researchers have reported the performance on LFW, here for fairly comparison and avoid the repeated work, we don't set another evaluating standard, in face verification problem, our database can be served as an external training data. Compared with training on LFW, our data set has following advantages:

1) Our data set contains much more face images which allow the user generate more face pairs for training. Consider the number of face images in our database, the underlying face pairs can be made is extremely huge.
2) For each identity in our data set, they have much more images than that in LFW, thus our data set is suitable for supervised learning algorithm, and it broaden the development of face verification algorithm. Previously, LFW is not suitable for supervised training, thus limited the algorithm development, most of the algorithms trained on LFW are semi-supervised learning algorithm [15][23][25].
3) Our data set contains more variations such as ethic, age and variation caused by makeup or cosmetic. In LFW, most of the people are European and American, just a small proportion of images come from Asian people, in contrast, our data set has a balance between Asian people and Western people. Additionally, the images in LFW has little age variations, most of the face images are collected from a certain short period, in our data set, a significant number of people have face images in different period such as teenagers and adult time. Finally, since these days, quite many people makeup everyday or cosmetic from time to time, our data set contains enough variations caused by cosmetic or cosmetic.

B. Face Identification

Our data set is intended for identification. In our data set, we have training sub set and test sub set. A critical aspect of our database is that for any given training testing split, the people in each subset are mutually exclusive.

In our data set, we fully support the training and test developing method, we organize our data into three Views, similar with that in LFW. View 1 is for algorithm development and general experiment before final evaluation. This might also be called a model selection or validation view. View 2, is used for performance reporting, should be used only for the final evaluation of a method. We need keep in mind that we should use the final test set as seldom as possible before reporting. Ideally, each test set should only be used once, i.e., when we report the final performance. View 3 is served a image

[2] For more details related to this performance, readers can email me

collection from the unknown identities when we use open set test to measure the performance.

View 1: Model selection and algorithm development. This view consists of two subsets of the database, one for training, and one for testing. The training set contains 400 identities and around 8000 images. The test set consists of 100 identities and 2000 images. To avoid the underlying over fitting problem, the people who appear in the training and testing sets are mutually exclusive. In the training set, developers can use any kind of method to training, for example, training with image pair data organized from those training images or training with label information for each face image. In the test set, however, for each identity, the face image registered in the gallery is fixed, and the remaining images in each identity is used as probe images. The main purpose of this view is that researchers can freely experiment with algorithms and parameter settings without worrying about overusing test data. To use this view, simply train an algorithm on the training set and test on the test set.

View 2: The second view of the data should be used only for performance reporting. In ideal case, it should only be used one time, as we all know, choosing the best parameter from multiple parameters, will bias results toward artificially high accuracy. The second view of the data consists of two subsets of the database. Once some parameters or algorithms have been selected (using View 1 of the database if desired), the performance of that algorithm can be measured using View 2. To report accuracy results on View 2, the experimenter should report the performance of a classifier on the two separate data set in a leave one out cross validation scheme. In each experiment, one subset can be used as a training set, with the other subset used for testing. It is critical for accuracy performance reporting that the final parameters of the classifier under each experiment be set using only the training data for that experiment. In other words, when we report the performance, each of the 2 experiments (both the training and testing phases) should be run completely independently of the other one, resulting in 2 separate classifiers (one for each test set). Also, as mentioned before, each time the test set can be only used for one time. A common violation for this rule is that to obtain the best performance on the test set, we training on the training set for multiple times, e.g., using SVM, and each time we test it on the test data to get the highest test performance [27], when performance reporting, we just report the best performance.

View 3: It contains face images from 200 identities, all these images has no overlap with the images in View 1 and View 2. This view can be used as the unknown identities if we measure the open set test performance, otherwise, we don't use the images in this view.

Correct use of training, validation, and testing sets is crucial for the accurate comparison of face recognition algorithms. In NIST Face Vendor Test, the competitors cannot access the test sets, even there is no publicly available training set, this is the most strict scenario. However, in general case as said in our paper, the researcher usually can access some training data, but the algorithm developers should not test their parameters of their algorithms to the test data before report the final performance.

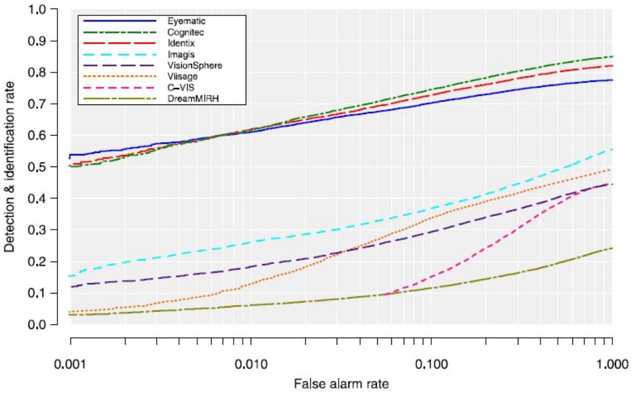

Fig. 2. Open Set Performance on ROC

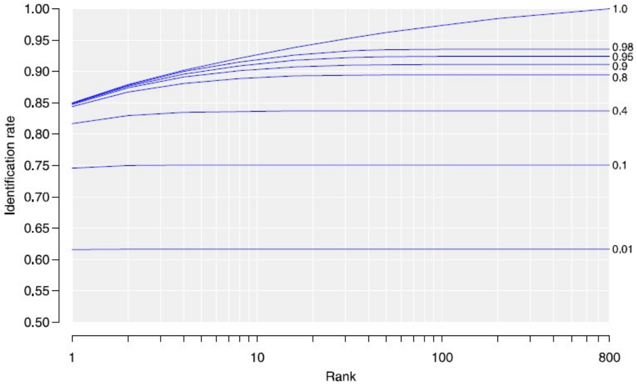

Fig. 3. Open Set Performance as a function of rank for eight FARs

C. Performance Reporting

Since our data set is intended for face identification, we use two performance measurements, one is accumulated curve performance, which might be proper for close set test. Another measurement considers the false alarm rate, which is more realistic in real face recognition applications. Each developer should report both these two performance since we find that two performance measurements have no direct relationship.

For open set identification, two figures must be reported, Fig.2 and Fig.3 can be referred as examples (Figures come from [17]). For close set identification, the CMC curve must be reported, as the example shown in Fig.4 (Figure comes from [17]).

In Fig. 2, the x axis is the false alarm rate while the y axis is the identification rate, remember that when calculate the identification rate,developers should not calculate the false rejected images or the false identified images. In Fig. 3, the x axis is the rank number while the y axis is the identification rate, and this figure focus on the rank N performance under different false alarm rates, here the developers should report the rank N performance under the following false alarm rate: 0.01, 0.05, 0.1, 0.3, 0.5. For more details related to these measurements, developers can refer to [17].

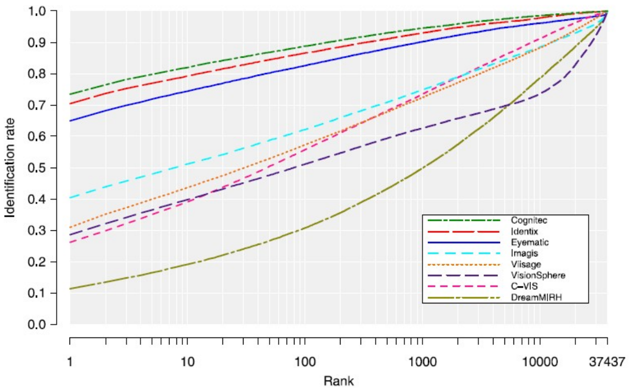

Fig. 4. Close Set Identification Performance on CMC

V. Conclusion

In this paper, a new database, namely, POS Faces in the Wild, is introduced. Our purpose for collecting this data set are as following: provide a large database of real world face images with enough variations to study the face identification problem; provide a large external training data set for face verification problem. By measuring performance on our data set, it allow the researcher compare their face identification performance in the same framework. We will release our data set for academic usage.[3]

Acknowledgment

1. The research was supported by the Implementation of Technologies for Identification, Behavior, and Location of Human based on Sensor Network Fusion Program through the Ministry of Trade, Industry and Energy (Grant Number: 10041629)

2. This work was supported by the MKE (The Ministry of Knowledge Economy), Korea,under the Core Technology Development for Breakthrough of Robot Vision Research support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2012-H1502-12-1002).

References

[1] A. M. Martinez and R. Benavente. The ar face database. Technical Report 24, Computer Vision Center, University of Barcelona, 1998.

[2] Ferdinando Samaria and Andy Harter. Parameterisation of a stochastic model for human face identi?cation. In Proceedings of the Second Workshop on Applications of Computer Vision, Sarasota, Florida, 1994.

[3] Anelia Angelova, Yaser Abu-Mostafa, and Pietro Perona. Pruning training sets for learning of object categories. In CVPR, volume 1, pages 495?501, 2005.

[4] Wen Gao, Bo Cao, Shiguang Shan, Delong Zhou, Xiaohua Zhang, and Debin Zhao. The CAS-PEAL large-scale Chinese face database and baseline evaluations. Technical Report JDL-TR-04-FR-001, Joint Research and Development Laboratory (China), 2004.

[5] P. Jonathon Phillips, Patrick J. Flynn, Todd Scruggs, Kevin Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the Face Recognition Grand Challenge. In CVPR, 2005.

[6] National Institute of Standards and Technology. The Color FERET Database. http://www.itl.nist.gov/iad/humanid/colorferet/home.html.2003.

[7] M. U. Ramos Sanchez, J. Matas, and J. Kittler. Statistical chromaticity models for lip tracking with B-splines. In International Conference on Audio and Video-Based Biometric Person Authentication, 1997.

[8] Craig I. Watson. Nist mugshot identification database. http://www.nist.gov/srd/nistsd18.htm, 1994.

[9] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. PAMI, 25(12):1615-1618, 2003.

[10] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class speci?c linear projections. IEEE Pattern Analysis and Machine Intelligence, 19(7), 1997.

[11] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(6):643-660, 2001.

[12] G. B. Huang, M. Ramesh, T. Berg, and E. L. Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. TR of University of Massachusetts, Amherst, Oct, 2007.

[13] H. S. Lee, S. Park, B. Kang, J. Shin, J. Y. Lee, H. M. Je, B. Jun, and D. Kim, The POSTECH Face Database (PF07) and Performance Evaluation. Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition, 2008.

[14] G. Hua, M. H. Yang, and Y. Ma, Introduction to the Special Section on Real-World Face Recognition, IEEE Transactions on PAMI, Vol.33, Issue.10, Oct, 2011.

[15] H. V. Nguyen, and L. Bai, Cosine Similarity Metric Learning for Face Verification, Lecture Notes in Computer Science, Vol. 6493, pp 709-720, 2011.

[16] B. Kang, J. Yoon, H. Wang, D. J. Kim, Affine Dense SURF, 2013 Conference of Samsung Techwins Research Center, Seoul, Korea, 2013.

[17] Handbook of Face Recognition, S.Z. Li and A.K. Jain, eds. Springer, 2005.

[18] O. Jesorsky, K. Kirchberg, and R. Frischolz. Robust face detection using the Hausdorff distance. In J. Bigun and F. Smeraldi, editors, Audio and Video Based Person Authentication, pages 90C95. Springer, 2001.

[19] Georgia Institute of Technology. The Georgia Tech Face Database. ftp://ftp.ee.gatech.edu/pub/users/hayes/facedb/.

[20] E. Marszalec, B. Martinkauppi, M. Soriano, and M. Pietikainen. A physics-based face database for color research. Journal of Electronic Imaging, 9(1):32C38, 2000.

[21] T. Ahonen, A. Hadid, M. Pietikinen, Face description with local binary patterns: Application to face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, Issue. 12, pp: 2037-2041, 2006.

[22] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, Information-theoretic metric learning, Proceedings of International Conference on Machine Learning(ICML), pp: 209-216, June, 2007.

[23] M. Guillaumin, J. Verbeek, and C. Schmid, Is that you? Metric Learning Approaches for Face Identification, International Conference on Computer Vision (ICCV), 2009.

[24] Y. M. Ying, and P. Li, Distance Metric Learning with Eigenvalue Optimization, Journal of Machine Learning Research, Vol. 13, pp: 1-26, 2012.

[25] C. Huang, S. H. Zhu, and K. Yu, Large Scale Strongly Supervised Ensemble Metric Learning, with Applications to Face Verification and Retrieval, NEC Technical Report TR115, 2011.

[26] D. Chen, X. D. Cao, L. W. Wang, F. Wen, and J. Sun, Bayesian Face Revisited: A Joint Formulation, European Conference on Computer Vision (ECCV), 2012.

[27] R. E. Fan, P. H. Chen, and C. J. Lin. Working set selection using the second order information for training SVM. Journal of Machine Learning Research 6, 1889-1918, 2005.

[3] Dataset can be obtained from www.mlcv.net.