Evaluation of a face recognition system performance's variation on a citizen passports database

Gabriel Lema*, Luis Di Martino†, Sebastián Berchesi‡, Alicia Fernández§, Federico Lecumberry¶ and Javier Preciozzi‖
Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay
Email: *gabolema@gmail.com, †lddm00@gmail.com, ‡sberchesi@gmail.com, §alicia@fing.edu.uy,
¶fefo@fing.edu.uy, ‖jprecio@fing.edu.uy

***Abstract*—Face recognition systems (FRS) have been widely studied and the performances reported are very high in the standard databases used for comparison. In this work we present a FRS that achieves state of the art results in these databases and show its performance's variation when tested in a field trial using a citizen identification database. To accomplish this, a set of experiments are proposed. These include increasing the size of the database, using subsets that include a time difference of one to ten years between the query samples and those enrolled in the system and finally using different subsets of the same database. Discussion on these experiments and conclusions are presented.**

I. Introduction

Face recognition systems (FRS) have achieved very high rates of success, with several methods reporting recognition rates of 98% and more [3], [15], [19], [20], [21]. These results were achieved using datasets in which images are acquired in controlled conditions such as FERET [14]. Such conditions are usually not met in real life situations, e.g. checkpoints at airports, where FRS are used for identification or authentication of identities. This fosters a great interest in understanding how big changes in illumination and pose as well as those introduced by the aging process affects the performance of FRS. Both issues have been addressed in the literature, a complete review of how aging affects the performance of a FRS and different approaches used in order to model the aging process can be found in [12]. Performance degradation due to changes in illumination, pose and occlusions is currently being tackled by using the *Labeled Faces in the Wild (LFW)* [7] database. In [5] the authors provide a good comparison of the performance obtained using both FERET and LFW databases.

In this work we present a FRS that achieves state of the art recognition rate on a controlled database as FERET, and we analyze its performance when faced against a more complex database. This database was obtained from a passport and identity card system; therefore factors such as pose, illumination, database size, small samples per class and aging come into play. We refer the reader to [10] and [13] for previous studies that performed a systematic evaluation of a verification system using a real dataset obtained from a passport system.

Considering the database properties is not possible to use statistical methods that need several samples per class. Thus, we focus our attention in methods that rely on only one sample per class. A vast number of techniques [3], [15], [19], [20], [21] have been presented based on one sample per class. These have proven to achieve recognition rates in the range of $97-99\%$, using small datasets with photos acquired under controlled conditions (as the FERET database). For example using micro patterns techniques [3] a recognition rate of 97% was achieved using the FERET *fb* probe set. Adding Gabor wavelets as part of the feature extraction procedure [21] has increased the recognition rate up to 98%. Later, a recognition rate of 99% has been achieved using a combination of local and global features for the face representation [15]. When faced to the aging problem, we found that there is not a definitive solution presented in the literature for methods based on one sample per class. The techniques [12], [17] are the ones that obtain the best results, and both create an aging model for each individual. These models are practically impossible to conceive in citizen identification because in these kind of databases there are usually no more than a couple of photos per person. Recently, big efforts are being done to overcome this difficulty and simulate the aging process by using a unique image of the subject in the gallery. A good example of such approach can be found in [8]. However, the proposed solution has not been tested yet on an automatic FRS.

The rest of the paper is organized as follows. In Section II we present the goals of the present work and in Section III the developed framework. In Sections IV and V we present the proposed experiments and the results, respectively. Finally, in Section VI we highlight the conclusions and propose directions for future work.

II. Goal of this work

The main goal of this work is to analyze how the performance of an automatic FRS is affected when using a dataset obtained from a real passport/ID issuance office.

In order to achieve this goal we first need a testing FRS. We develop a FRS by using state of the art techniques in each of the stages involved in the recognition task. Then, we study the performance of the constructed system in a well controlled conditions database (FERET). We continue with the evaluation of the performance of the FRS using the real passport/ID

database. This dataset presents several features introduced before: changes in illumination, gesture and expression, few number of samples per class, big size of the dataset and age difference between samples. We outline a set of experiments to investigate each one of the variables that affects a FRS.

III. Developed framework

As in many other biometric and pattern recognition systems we divided the FRS into three modules: preprocessing, feature extraction and matching. This framework design permitted us to focus on each module individually and easily change the used algorithms in each stage seeking for the best performance in each particular task.

1) Preprocessing module: The preprocessing module automatically validates the face in an input image, discards all unnecessary information, and transforms the face image in order to comply with a predefined standard. This standard was defined by fixing the size of the face and the eyes positions. The preprocessing module is very important in an automatic FRS, since it is the one that can estimate the quality of the captured face image. Based on the quality of the face image the system could ask for a new acquisition of the face or mark the final result as not totally reliable. This approximation is used in [5], where a "sample pose quality index" and "sample illumination quality index" are computed and used in the process of evaluation of sample distortion and normalization. Additionally, the performance of the preprocessing stage affects the final result when local features are used, since the more accurate the eyes localization is, the better the registration of the face in the input image will be.

Different approaches could be used in order to automatically find eyes positions. Some techniques act locally by searching for the eyes in a moving window over the image. A good example is the implementation of the eye detector in the OpenCV library [2] based in the work [16].

Local methods are not well suited in the case one or both eyes are closed or missing in the face in the input image. This situation is very rare in a testing database as FERET, but feasible in a citizens database. A solution to this problem is to use a global approach that returns the eyes positions as well as others landmarks in the face. A popular technique in this category is Active Shape Models (ASM) [4]. There are several open source implementations of the ASM method: OpenASM [1], ASMLibrary [18] and STASM [11]. We choose STASM because it is easy to integrate, already includes a trained model and performs better than OpenASM and ASM-Library (a comparison between different implementations of ASM is performed in [5]).

Once the eyes positions are determined, the image is normalized using a transformation that takes the eyes to the predefined positions. Fig. 1 shows an example of the input image and the resultant image.

2) Feature extraction module: The feature extraction module receives a normalized and preprocessed image and returns a feature vector. This feature vector would, ideally, uniquely represent the person in the input image.

We used the method presented on the article [3] for various reasons: it is very simple to implement, fast in processing an input image, and provides very good results. Additionally, it does not require a statistical training, making it suitable to the case in which we have only few (or even one) sample per person in the gallery dataset. A complete review of more complex LBP based techniques can be found in [6].

In this approach, the face image is divided into several regions from which the LBP feature histograms are extracted and concatenated as shown in Fig. 2. For a more detailed description of its operation we refer the reader to the original article [3].

3) Matching module: The matching module receives a feature vector, compares it against the feature vectors of the enrolled subjects, and returns an identity. As a consequence of the selected feature extraction technique, the feature vectors are composed by the concatenation of several histograms where each histogram corresponds to a particular face region. Therefore, it is possible to use any distance between histograms to perform the comparison. We used the Chi-Square distance, since it has been widely used when working with LBP histograms. Given a face q_i in the input image and a face g_j in the gallery dataset, the distance $D(q_i, g_j)$ between them is computed as shown in Equation 1

$$\begin{aligned} D(q_i, g_j) &= \sum_{n=1}^{N_p} w_n \chi^2(q_{i,n}, g_{j,n}) \\ &= \frac{1}{2} \sum_{n=1}^{N_p} w_n \left(\sum_{m=1}^{N_b} \frac{(q_{i,n,m} - g_{j,n,m})^2}{(q_{i,n,m} + g_{j,n,m})} \right) \end{aligned} \quad (1)$$

Where N_b is the number of bins each histogram has, N_p is the number of patches (or histograms) each picture has, and w_n is the weight we give to each patch.

Since each patch represents a region of the face, we can assign a high weight to those which play a more important role in face recognition, and a small weight to those which do not. Fig. 3 shows an example of these weights, where the darkest the region, the lowest the weight. Once the distances between the extracted features of the query face image and all the features vectors in the gallery are computed, the matching

Fig. 1. Original and normalized images

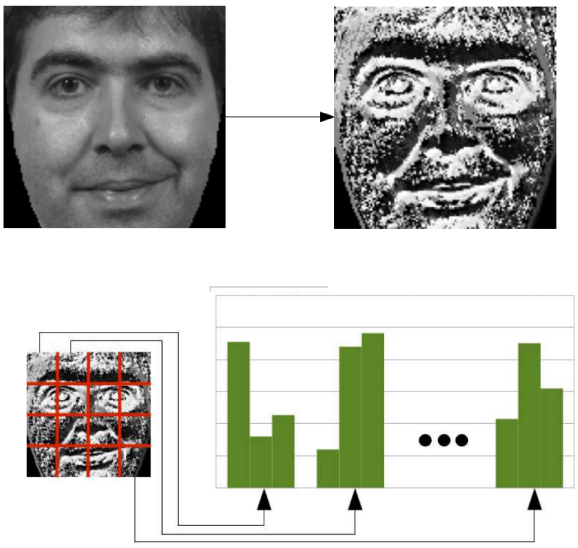

Fig. 2. LBP processed image (top) and generated histograms (bottom).

is done using a nearest neighbor approach. The query subject is assigned the identity of the enrolled person with the lowest distance to the person in the input image.

IV. Experimental setup

One of the main goals of the present work is to analyze the degradation on performance of the presented algorithms when faced to a citizen passports/IDs database. In this section we introduce the performed experiments using the framework presented in Section III. These experiments can be grouped into two different categories:

- Experiments on standard databases. A set of experiments were performed on the Color *FERET* dataset using the query dataset f_b to recognize the identities enrolled in the gallery f_a. This first experiment allows us to validate the proposed framework, showing that it achieves state-of-the-art results when working under controlled conditions.
  In these datasets the eyes positions were manually marked and coordinates are provided. Therefore, this first test also allows us to measure the degradation of the performance due to errors on face registration when using an automatic eyes finder.
- Experiments on *DNIC* dataset. This is the core of the present work. In this set of experiments, we study how the performance of the proposed FRS is degraded when faced to a dataset obtained from an in-production database containing citizens passports and IDs.

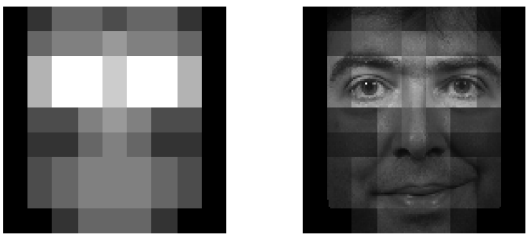

Fig. 3. Example of weights assigned to the face patches

A. Used databases

1) FERET: The FERET database was created as part of the program Face Recognition Technology carried on in the years $1993 - 1997$ [14]. This database has become very popular in the area of face recognition and is commonly used as a benchmark. It contains a gallery f_a containing 1010 people and four standards test sets: f_b, f_c, dup_1 and dup_2. In this work we use the f_b set that includes images taken on the same day that the ones taken for the gallery including only differences in expression.

2) DNIC: *DNIC* (Dirección Nacional de Identificación Civil) is the Uruguayan government organization responsible for the emission of ID cards and passports. It has a civil identification system, and a database with more than 3 million identities [1]. The images have some desired properties as neutral pose, neutral background and no lens among others. Despite this, large variations in illumination and expressions as well as aging are to be expected considering the huge amount of images acquired each day in the different offices throughout the country.

B. Proposed evaluations

In this work we perform a closed-set identification; for each sample in the query database there is only one correspoding sample in the gallery and viceversa. The performance is measured as follows [9]: for each sample in the query set q_i we compute its distance to each of the samples in the gallery and sort them from lowest to highest. Let $\tilde{g}_k$ be the sample from the gallery that has the same identity as q_i. The probe q_i has rank n if $D(q_i, \tilde{g}_k)$ is the n^{th} smallest distance. The recognition rate for rank n is the fraction of query samples that have rank n or lower. When the rank is not specified, the reader should assume that it is rank 1.

In order to evaluate the performance of the system using the presented datasets we propose four different experiments:

- Impact of production environment: comparison between FERET and DNIC. We study the performance of the presented FRS with both FERET and a subset of the DNIC database. Both databases contain the same amount of people in gallery and query sets and no time difference between samples. The main difference between the FERET and DNIC dataset is that the latter is obtained from an on-production environment where the acquisition is not strictly supervised as in a lab acquisition scenario. This implies that the images may have changes in illumination and face expressions among others.
- Variability of the results. In this experiment we analyze the variance in the performance of the system as different datasets are used. Generally this type of

[1] Because of the local laws that protect privacy, samples of this database cannot be shown in this work.

analysis is not done when reporting the performance of a FRS. We use ten query and gallery datasets, each one containing one thousand people with no time difference between them.

- Impact of dataset size. In this experiment we test how the dataset's size affects the performance of the FRS. To accomplish this, we use a subset of the DNIC database containing ten thousand people with no time difference between samples in gallery and query sets. As detailed previously, conditions as illumination, pose and face expressions are semi-controlled.
- Impact of age between samples. With this experiment we not only try to analyze the impact of aging in the FRS performance, but also its variations across the time difference gap. We use ten query and gallery datasets, each one containing two hundred people over 18 years. These sets contain face images with difference of one to ten years with respect to the ones in the gallery.

V. Experimental results

A. Comparison with a standard dataset

We start our set of experiments on FERET dataset. Most of the work presented in the literature use preprocessed images with manual eye localization. Nevertheless, in most practical situations, like in the DNIC dataset, it is not feasible to mark eye positions manually. To analyze the effect of automatic detection of eye positions in the performance of the system we compute the results using both manual, and automatic eye detection. Obtained results are shown in Fig. 4. It can be seen that even under this well controlled situation the accuracy of the method drops down as far as 3%, which leads us to our first preliminary conclusion: eye localization has a direct impact on the overall performance. This can be explained because the eyes positions are the used reference when the image is normalized. Additionally, when using features like LBP the normalized image is divided in fixed patches. Thus, differences in the localization of eyes positions affects the registration and extracted features of the face.

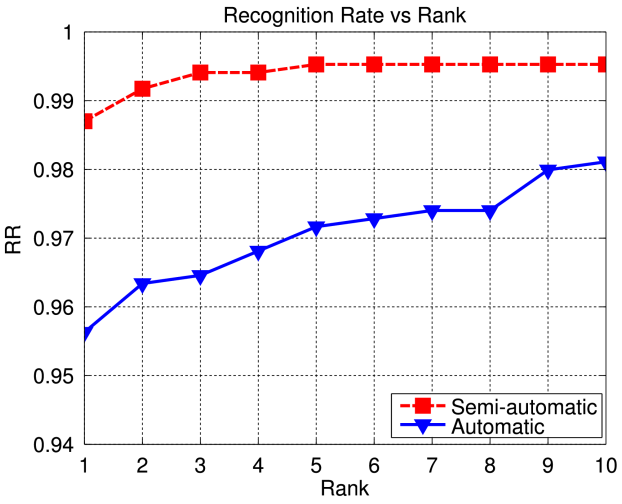

Fig. 4. RR vs Rank for manual and automatic localization of eye positions in set fb of FERET database

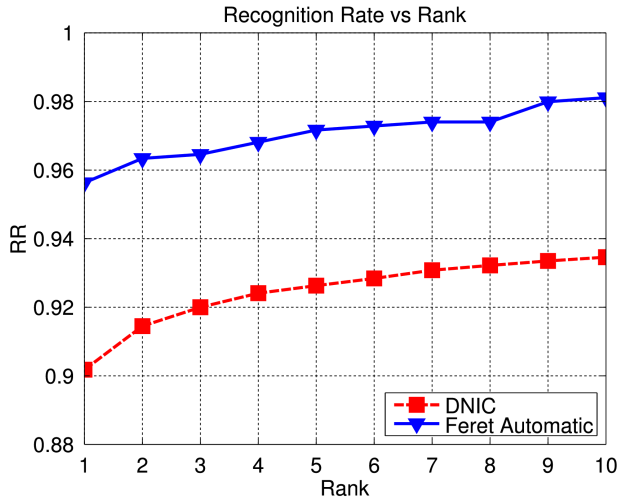

Fig. 5. RR vs Rank, comparison between Feret and DNIC databases.

B. Application to the DNIC dataset

1) Comparison with FERET dataset: We start our experiments on DNIC dataset with a subset selection as similar as possible to the FERET fb probe set used on the previous experiment. This first subset is composed of a gallery and a query dataset of 1000 images taken on the same day, with people over 18 years old. The results are shown on Fig. 5. The main difference between this dataset and the FERET, is the production environment conditions in which the former was acquired. It can be seen in the figure that the system behaves similarly on both databases, but with a 5% drop on the DNIC database. Even though it is very difficult to find the specific cause of this degradation, we believe it should be attributed to the variations on pose, illumination or expressions present in the DNIC database.

2) Consistency between databases: In this experiment, we discuss a generally underestimated issue when reporting biometrics results. It is usual when testing a biometric system, to have access to a bounded and unique database or a set of databases very dissimilar from each other. This prevents us from performing a statistical study of how the FRS will perform under different scenarios. By working with the DNIC database we have access to different subsets of the same database, thereby allowing us to study the variance in performance as we vary the dataset.

To conduct this experiment, we used a gallery and query set of 10.000 people. We then splitted them in 10 different query sets and gallery, so each query set has its corresponding gallery, and compute the performance of the system with each of them. In Fig. 6 the obtained recognition rate mean and 2σ confidence interval are shown in solid and dotted lines respectively. We can see that the confidence interval is around 2%, and therefore a positive variation of the recognition rate of $\pm 2\%$ should not be reported as an improvement since it would be on the variance range of the results.

3) Large dataset: In this experiment we analyze the performance of the system when faced with a large database. There-

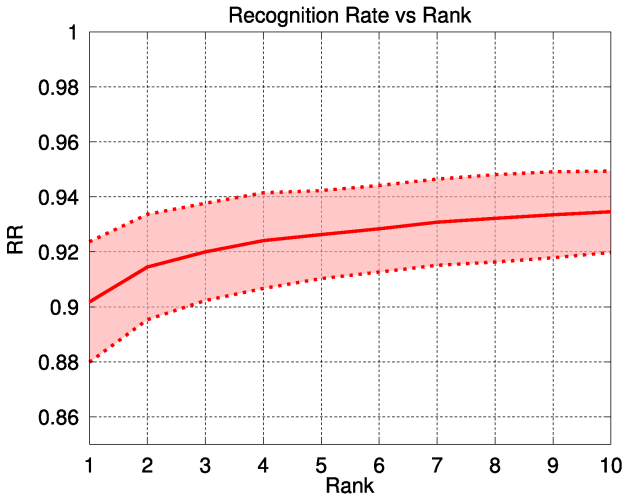

Fig. 6. RR vs Rank in DNIC database, mean and 2σ confidence interval plotted.

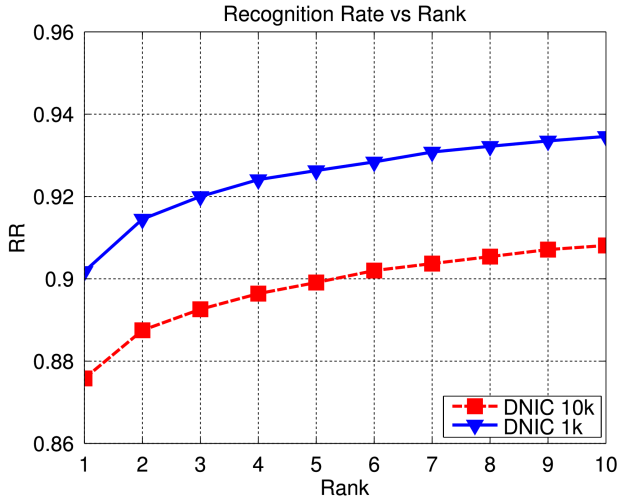

Fig. 7. RR vs Rank, comparison between DNIC datasets containing $1k$ and $10k$ people.

fore we use the gallery and query set of 10.000 faces, instead of 1000. The results are reported using the same methods and configurations as before and are shown on Fig. 7. As expected, when we increase the number of people in the database, since we are including more people that might present similar features in their faces, we increase the probability of finding missmatches. However we were surprised to find that there was no major decrease in the performance, but only a 3%. This result is very promising when analyzing the possibility of using a FRS in a passports/IDs issuance office where they handle a large number of people. It would be interesting to see if the same drop in the performance is obtained with a ten-fold increase in the number of people in the database.

4) Aging: It is well known that one of the problems that affects more the performance of FRS is aging. In this final experiment we will examine the relation between age difference between images and the overall performance of the system. To accomplish this, we use 10 sets of images that have age difference with respect to the images in gallery

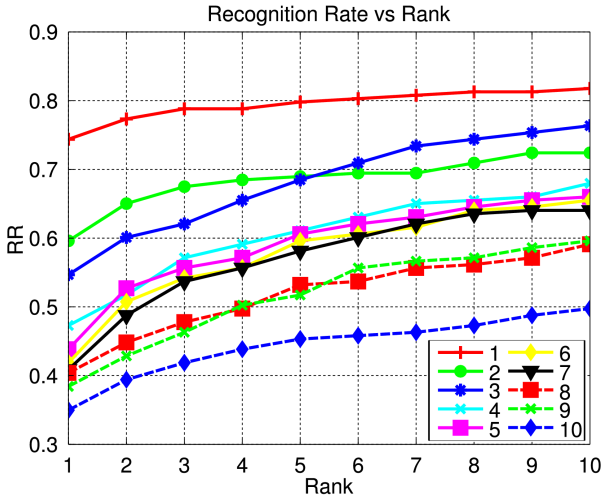

Fig. 8. RR vs Rank in DNIC database, time difference between images. Each color represents the years between samples. It is clear the performance degradation.

ranging from 1 to 10 years. Each of these sets contains 200 people older than 18 years. Fig. 8 shows the results obtained in this experiment. These do not show the same behavior as the ones obtained on [10], since they report a stabilization on the performance starting at a gap of 4 years and the results on this experiment shows that the performance continue to degrade along the years. Further experiments must be carried out to confirm or reject this hypothesis.

VI. Conclusions and Future Work

In this work, we present an analysis of the performance variation of a FRS when faced to a real database composed by passports/IDs images. We have highlighted most of the well known problems in face recognition and we have analyzed their impact on the performance of a state-of-the-art FRS.

We first showed that eye localization has a direct impact on the performance of the system when local features are used. This is an important result since many work is devoted to obtain better facial recognition methods using manually marked eyes positions, when it is clear that this approach is not valid in real applications where eyes coordinates are not provided. Secondly, we showed that when we tested our FRS with a dataset obtained in an on-production environment, the performance drops dramatically compared with a controlled dataset. Clearly, work must be done to understand the differences between both datasets in order to propose systems with better performance on real databases. We also showed that a FRS performance's drops when we use a larger database, but not too much. To better validate this, further experiments should be conducted using higher orders of magnitudes, e.g. 100.000 people.

One of the main contributions of the present work is that we analyze what is the variance of the results when we perform several independent statistical experiments. For our FRS, we conclude that the confidence interval is around 2%.

Regarding the aging process, we conclude that it remains as one of the biggest problem that FRS faces. Our evaluation

shows that the proposed framework, that achieves a high performance when used with images taken in the same day, is highly degraded when age difference is present between the images in the query and gallery sets. Additionally we found that when the time difference gap is bigger than three years the identification task becomes very hard. We should obtain better results with methods that build an aging model of each subject, but these methods are not applicable when using a dataset like the DNIC database in which there are a few samples per person. Last but not least, we should look for a better set of weights in the matching module, that would enhance our system's performance. This could be done using FLD as in [15].

ACKNOWLEDGMENT

The authors would like to thank the "Agencia Nacional de Investigación y Desarrollo" (*ANII*) for partially support this work and also the "Dirección Nacional de Identificación Civil" (*DNIC*) for providing access to their valuable database.

REFERENCES

[1] An asm implementation by c++ using opencv 2. https://code.google.com/p/asmlib-opencv/. Accessed: 2014-06-21.

[2] Opencv (open source computer vision). http://opencv.org/. Accessed: 2014-06-21.

[3] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037 –2041, dec. 2006.

[4] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38 – 59, 1995.

[5] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler. Robust face recognition for uncontrolled pose and illumination changes. *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, 43(1):149–163, Jan 2013.

[6] D. Huang, C. Shan, M. Ardabilian, W. Yunhong, and C. Liming. Local Binary Patterns and Its Application to Facial Image Analysis: A Survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 41(6):765–781, 2011.

[7] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[8] S. M. S. I. Kemelmacher-Shlizerman, S. Suwajanakorn. Illumination-aware age progression. In *CVPR*, 2014.

[9] S. Z. Li and A. K. Jain, editors. *Handbook of Face Recognition, 2nd Edition*. Springer, 2011.

[10] H. Ling, S. Soatto, N. Ramanathan, and D. Jacobs. A study of face recognition as people age. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1 –8, oct. 2007.

[11] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. *European Conference on Computer Vision (ECCV)*, 2008.

[12] U. Park, Y. Tong, and A. Jain. Age-invariant face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):947 –954, may 2010.

[13] P. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and M. Bone. Face recognition vendor test 2002. In *Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on*, pages 44–, Oct 2003.

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

[15] Y. Su, S. Shan, X. Chen, and W. Gao. Hierarchical ensemble of global and local classifiers for face recognition. *Image Processing, IEEE Transactions on*, 18(8):1885 –1896, aug. 2009.

[16] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511–I–518 vol.1, 2001.

[17] J. Wang, Y. Shang, G. Su, and X. Lin. Age simulation for face recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 913 –916, 0-0 2006.

[18] Y. Wei. Research on facial expression recognition and synthesis. *Master Thesis, Department of Computer Science and Technology, Nanjing University*, Feb 2009. http://code.google.com/p/asmlibrary.

[19] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):775 –779, jul 1997.

[20] B. Zhang, Y. Gao, S. Zhao, and J. Liu. Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor. *Image Processing, IEEE Transactions on*, 19(2):533–544, Feb 2010.

[21] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 786 – 791 Vol. 1, oct. 2005.