Can Humans and Automatic Algorithms Recognize Look-Alike Faces?

Hemank Lamba, Ankit Sarkar, Mayank Vatsa and Richa Singh
IIIT Delhi
New Delhi, India
{mayank, rsingh}@iiitd.ac.in

Abstract

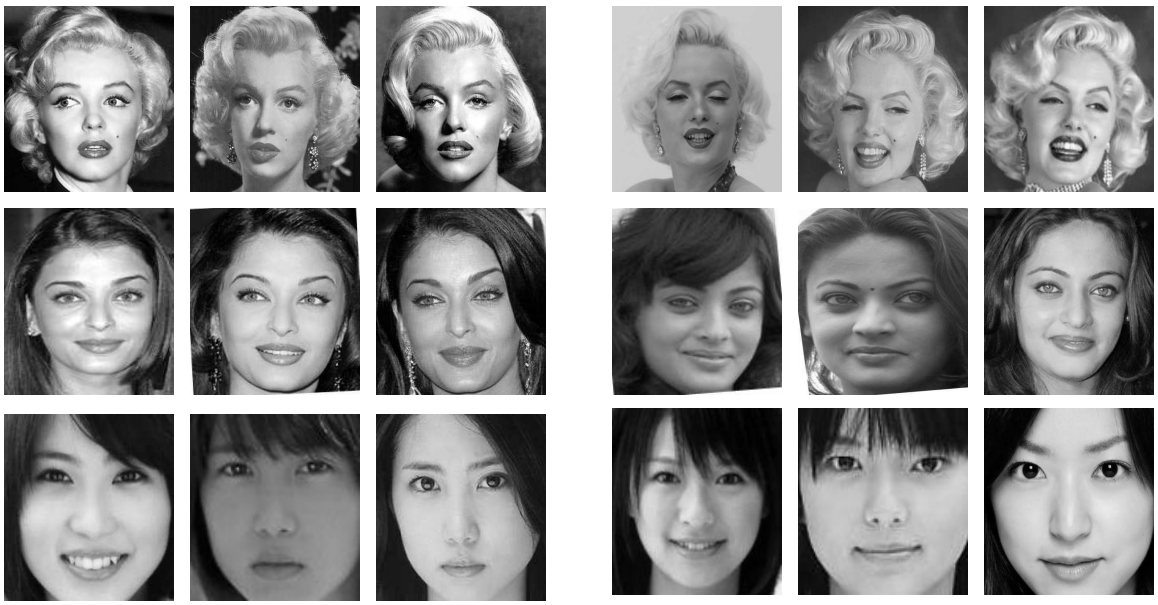
One of the major challenges of face recognition is to design a feature extractor that reduces the intra-class variations and increases the inter-class variations. The feature extraction algorithm has to be robust enough to extract similar features for a particular class despite variations in quality, pose, illumination, expression, aging and disguise. The problem is exacerbated when there are two individuals with lower inter-class variations, i.e., look-alikes. In such cases, the intra-class similarity is higher than the inter-class variation for these two individuals. This research explores the problem of look-alikes faces and their effect on human performance and automatic face recognition algorithms. There is two fold contribution in this research: firstly, we analyze human recognition capabilities for look-alike appearances and secondly, compare it with automatic face recognition algorithms. In our analysis, we observe that neither humans nor automatic face recognition algorithms are efficient for the challenge of look-alikes.

1. INTRODUCTION

Humans effortlessly process information obtained from multiple sensory inputs and have the ability to recognize individuals even with limited correlation information, redundant information, or when certain features appear partially hidden, camouflaged or disguised. To recognize an individual, the visual cortex exploits spatial correlations by processing overlapping information extracted at global and local levels and effectively combines them to make a decision. The information is gathered using a set of inherent spatial filters that accurately detects any change in orientation, color, spatial frequency, texture, motion, and other pertinent features. For several years, many researchers have been motivated in developing algorithms to emulate the near perfect face recognition capability of human mind. However, human face is not a rigid object and can have different variations due to inter-personal or intra-personal transformations. Inter-personal variations can be attributed to changes in race or genetics, while intra-personal variations can be attributed to changes in expression, pose, illumination, aging, hair, cosmetics, and facial accessories. These inter and intra-personal variations can be easily deceived by *look-alike* faces or using *disguise* tools. In this paper, we specifically undertake the challenge of face recognition with look-alike variations.

As shown in Fig. 1, face recognition algorithms may fail when they are encountered with similar looking faces, or as we may say, *look-alikes*. Most of the existing automatic face recognition algorithms are based on appearance, feature and/or texture based models to identify individuals. Nonetheless, these algorithms will obviously fail in the context of look-alikes because both the individuals (look-alikes) will have near identical appearance, feature and, maybe, texture. This assertion is based on the study by Kosmerlj *et. al.* [1]. In this study, an experiment was conducted to estimate percentage of Norwegian people having one or more look-alikes. This study concluded that face recognition technology may not be adequate for identity verification in large scale applications, particularly under the presence of look-alikes. In cognitive science, similar topic has been discussed from a different point of view - other race effect on face recognition. In other race effect, an individual may not be able to correctly recognize faces from other races and believes that faces from other race look alike. Carpenter [2] suggests that it is not that the people cannot perceive subtle differences among those who belong to other racial groups. It is rather that they lay more emphasis on recognizing the race of the person whether he is black, asian or white and they do not explore a person's distinguishing features. It is a developed hypothesis that people recognize faces of their own race more accurately than faces of other races. The *contact* hypothesis suggests that other race effect occurs as a result of greater experience we have with own- versus other-race faces [3].

Many forensic and law enforcement applications have to deal with this important challenge. However, this covariate has not received much of attention from the research community. The challenge of look-alikes is studied by the



(a) (b)

Figure 1. Examples of look-alikes: (a) genuine and (b) look-alikes of respective individuals in (a).

cognitive scientists but no proper evaluation has been performed for automatic algorithms. It is our belief that we need to analyze the performance of human recognition capabilities under this covariate and compare it with automatic algorithms. Contribution of this paper is therefore two fold: (1) analyze human recognition capabilities for look-alike appearances and (2) compare it with automatic face recognition algorithm. For automatic face recognition algorithm, we use algorithms based on kernel subspace analysis and their linear counterparts as well as texture descriptors based algorithms.

2. Human Recognition Capability for Look-alike Faces

To the best of our knowledge, there is no experimental study that evaluates human capabilities as well as automated algorithms to recognize look-alike face images. It is our opinion that, such an evaluation is important in designing newer and better algorithms that can recognize images with this covariate. To evaluate the performance of human recognition capabilities, we have prepared a look-alike database.

2.1. Look-alike Face Database

It is extremely difficult to prepare such an database. However, different web-sites presents several look-alike cases, specially for celebrities and known individuals. We have collected these cases and prepared the *look-alike database*. This database consists of images pertaining to 50 well known personalities (from western, eastern and asian origin) and their look-alikes. Each subject/class have five genuine images (total 50×5 genuine cases) and five look-alike images (total 50×5 look-alikes). While collecting these images, it was ensured that the images for every class do not have any major variation in pose/illumination. It was also made sure that images did not differ in the amount of makeup and other accessories.

2.2. Human Evaluation Protocol

For human evaluation, group of 50 volunteers were requested to participate. Here are some statistics about the human volunteers:

- Age Variation : 10 to 57 years
- Gender Variation: 20 female 30 male
- General Background: Majority of undergraduate students as well as children of age around 10-12 years and housewives.

The volunteers were shown different face images from the *look-alike database* and they had to recognize and find genuine pairs. They were shown easy pairs as well as difficult pairs for recognition purposes. To find easy and difficult pairs we used PCA approach [1]. Using these pair, the volunteers were asked to rate the similarity between two images on a scale of 1 to 5.

[1] we first applied PCA on the database. Using the PCA scores, we identified easiest and the most difficult pair of images. The *most difficult* pair was selected as one which had the least similarity score within the class and the *easiest* pair was selected as the one which had the highest similarity score.

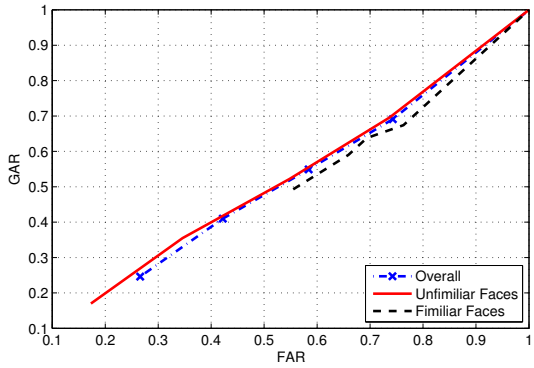


Figure 2. ROC plots for human analysis on the look-alike database.

5 = Sure they are the same
4 = Think that they are the same
3 = Do not know
2 = Think they are not the same
1 = Sure they are not the same

Besides this, for every pair, volunteers were given a specific time of 10 seconds to identify and rate the images. This was done because in real world scenario such as border control, normally a human evaluator has about 10 seconds to look at the individual's face and document. To better analyze the results, we also asked whether the volunteers to mention if they had known the given pair from past (familiar vs. unfamiliar faces). Finally, volunteers were asked to mention what specific features they used in recognizing faces.

2.3. Results and Analysis

The responses were compiled and the receiver operating characteristics (ROC) curves were generated. Fig. 2 shows the results of human responses. Key results are summarized below:

- The ROC curves show that human verification accuracy for look-alike database is very low. However, we observed that the volunteers performed better on classes with western and asian origins compared to Eastern origin.

- For some easy cases, volunteers easily performed correct verification whereas for complex cases, most volunteers were not able to perform correct verification. As shown in Fig. 3, first image pair was considered as easy and last pair was most complex where volunteers made the mistake.

- In our evaluation, interestingly, humans performed better on male classes compared to female classes. Based on the responses, an attempt was made to analyze this effect but we were not able to reach to any conclusions.

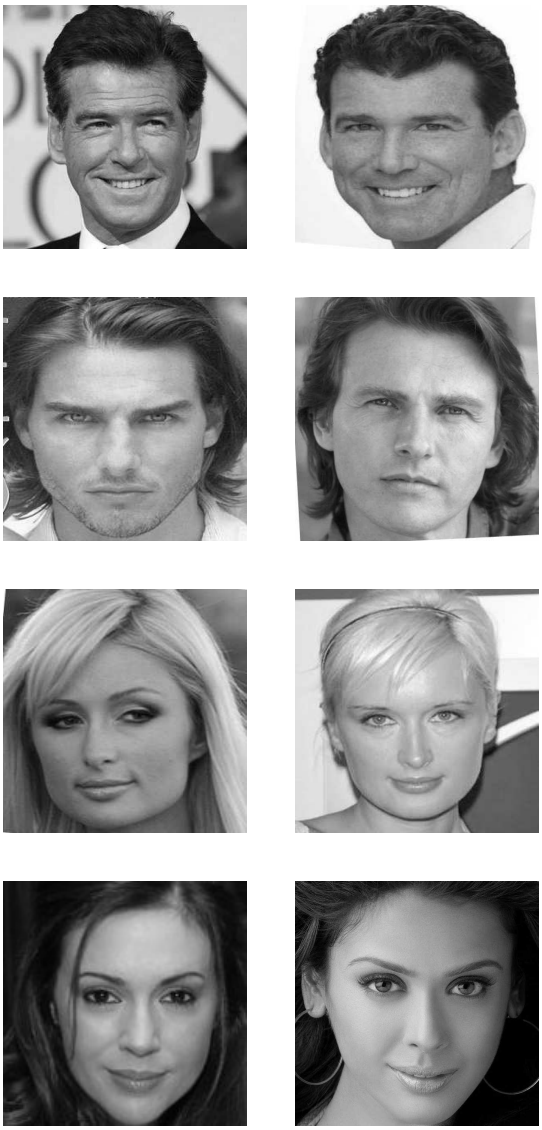


Figure 3. Examples of image pairs - images of two different individuals (look-alikes) - that are shown to the volunteers. These images are ordered in increasing order of complexity, i.e. first pair is easy and last pair is difficult.

Further, an attempt was made to analyze the effect of age of the volunteer on the verification performance. However, the sample space was very small to reach to any conclusions.

- In the experiments, *other race effect* did not affect the human performance. However, we observed that western origin is easier to match compared to other races.

- Since we requested the volunteers to mention about the (un)familiarity with the image pairs, we analyzed the effect of familiar vs. unfamiliar face recognition in humans. Generally, it is assumed that humans are

very good at familiar face recognition. However, using look-alike face database, ROC curves in Fig. 2 suggest that there is no significant difference between unfamiliar and familiar face recognition performance. This is an interesting result and, to the best of our knowledge, is not observed elsewhere.

- It was also noted that out of the 979 responses in human evaluation, timeout occurred only 37 times. A timeout means that the human was unable to judge the similarity within the allotted 10 seconds. However, increasing time did not help much in increasing the accuracy.

- In most of the responses, we observed that face shape, eyes, nose and lips play important role in making a decision. However, few responses also suggested that overall face appearance was discriminating.

We next compare the human performance with automatic face recognition performance.

3. Automatic Face Recognition Evaluation on Look-alike Database

Generally, face verification algorithms can be classified into four categories: geometry based, subspace based, texture descriptor based, and 3D approaches. In this research, we use subspace based and texture descriptor based algorithms for performance analysis on look-alike database.

3.1. Subspace Analysis Approach based Automatic Face Recognition Evaluation

Among various techniques, appearance based approaches have received major attention. These algorithms mostly use subspace analysis methods to address pose, expression, and illumination covariates. Examples of these algorithms include Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Independent Component Analysis (ICA). According to Vapnik-Chervonenkis theory, mappings from lower dimensional space (input space) to higher dimensional space, in general, provides increased classification capabilities [4]. However, increasing the dimensions can increase computational complexity. To overcome this issue and still get the benefits of higher dimensional mapping, kernel tricks are used. In face recognition, past research had shown that manifolds can be discriminating but with kernel tricks, discrimination capability of these subspace analysis approaches can be further enhanced. Therefore, researchers have introduced the use of kernel approach to subspace analysis and proposed kernel subspace analysis methods such as Kernel PCA (KPCA), Kernel LDA (KLDA), and Kernel ICA (KICA).

In this research, we use kernel subspace analysis methods and their linear counterparts for comparing the performance with human evaluation. Let $\mathbf{S} = [s_1, \cdots, s_N]$ be the d-dimensional vector where N is the number of elements in the vector. In face recognition, size of d is very large therefore, it is desired to find $b < d$ basis vectors such that

$$\mathbf{V} = \mathbf{W}^T\mathbf{S} \tag{1}$$

Here, $\mathbf{W} = [w_1, \cdots, w_b] \in R^d$. This representation is the basis of different subspace analysis methods. For example, in PCA, we try to minimize $||\mathbf{WV} - \mathbf{S}||$ using a scatter matrix $\mathbf{M} = \mathbf{SS^T}$. Detailed analysis of PCA, LDA, and ICA can be found in [5], [6].

In the non-linear kernel version of subspace approaches, the input data is mapped into a high dimensional feature space using kernel function where we can get better non-linear features. Conceptually, non-linear methods can capture higher-order statistics for better feature extraction. Mathematically, a non-linear mapping is defined as: $\phi(\cdot) : R^d \rightarrow R^h$, i.e. from a d dimensional R^d to a higher dimensional space R^h. In other words, every s_i in $\mathbf{S}$ is mapped to a point $\phi(s_i)$ in higher dimensional space. Now, we apply linear subspace method on the mapped feature space. Here ϕ can be any kernel function such as:

Linear kernel:

$$K(s_i, s_j) = s_i^T s_j \tag{2}$$

Polynomial kernel:

$$K(s_i, s_j) = (\gamma s_i^T s_j + r)^d, \qquad \gamma, r > 0 \tag{3}$$

RBF kernel:

$$K(s_i, s_j) = exp(-\gamma||s_i - s_j||^2), \qquad \gamma > 0 \tag{4}$$

Here, r and γ are kernel parameters. To understand little more, for N points in d-dimensions, we apply any one of the above kernels and then apply PCA in higher dimensional feature space. Note that, PCA will provide up to d principal components whereas N components will be computed by KPCA. More details of kernel approach for appearance based algorithms can be found in [4].

Before we test the algorithms on look-alike database, we first train and test the subspace analysis based algorithms on publicly available databases. This training-testing experiments on PCA-KPCA, ICA-KICA, and LDA-KLDA are performed using a large database with different challenging variations on pose, expression and illumination. We combined images from different face databases to create a non-homogeneous combined face database of 600 subjects. The AR face database[2] contains face images with varying illumination and accessories. The CMU-PIE [7] database

[2] http://cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html

Table 1. Composition of the non-homogeneous combined face database.

Face Database	Number of Classes (subjects)
AR	120
CMU - PIE	65
Notre Dame	315
Equinox	100
Total	600

Table 2. Verification accuracy of different kernels at 0.1% FAR for KPCA, KICA and KLDA.

Subspace Approach	Verification Accuracy (%)		
	Linear	Polynomial	RBF
KPCA	74.3	75.9	**77.2**
KICA	69.8	**71.6**	71.5
KLDA	76.5	77.7	**78.8**

contains images with variation in pose, illumination and facial expressions. The Notre Dame face database [8] comprises images with different lighting and facial expressions over a period of one year. The Equinox database[3] has images captured under different illumination conditions with accessories and expressions. Table 1 lists the databases used and the number of classes selected from the individual databases. Total number of images in the combined database is over 10,000 pertaining to 600 subjects. We divided the images into two sets: (1) training dataset and (2) gallery-probe test dataset. Training dataset is used to train the individual algorithms. It comprises images pertaining to 40% subjects (i.e. over 4000 images). Gallery-probe test dataset is used to evaluate the performance using remaining 60% subjects (i.e. over 6,000 image). Note that, in the experiments, training and testing data are not overlapping. This means that all the individuals in testing data are unseen or are not present in the training data. This train-test partitioning is repeated ten times and verification accuracies are reported at 0.1% FAR.

The optimal parameters for the kernels in non-linear subspace learning approaches (i.e. KPCA, KLDA, KICA) are obtained empirically by computing the verification accuracy for different combination of parameters. Table 2 shows the results obtained with optimal kernel parameters.

For KPCA, optimal parameter for the RBF kernel is $\gamma = 4$. For KICA, polynomial kernel provides the best results with $r = 1$ and $\gamma = 2$. Finally, RBF kernel with $\gamma = 6$ gives the optimal results for KLDA. The results in Table 2 show that for all three subspace analysis approaches, non-linear kernels provide higher verification performance compared to the linear kernel. This is because biometric match data is non-linearly distributed and hence non-linear kernels provide better classification. Using the optimal kernel parameters, Table 3 shows the comparative results between (a) PCA and KPCA, (b) FDA and KFDA, and (c) ICA and KICA. It has been found that non-linear kernel algorithms can better encode the facial features compared to their linear counterparts. Among linear methods, LDA provides best accuracy of 73% whereas KLDA gives the best results of 78.8% among the non-linear appearance based approaches.

[3] http://www.equinoxsensors.com/products/HID.html

Table 3. Verification accuracy of subspace based face verification algorithms on combined face database.

Algorithm	Verification Accuracy (%)
PCA	61.4
KPCA	77.2
ICA	62.7
KICA	71.6
LDA	73.0
KLDA	78.8

After training-testing on the non-homogeneous combined face database, trained algorithms are evaluated using the look-alike face database. In the verification experiments, we computed ROC curves for both kernel based algorithms and linear subspace algorithms. Figs. 6 and 5 show the results on the look-alike face database. These results clearly show that *look-alike* is a major challenge for face recognition. Equal error rates (ERR) for these automatic algorithms are above 50% which is not better than simple coin tossing. This is mainly because with look-alike face images, subspace analysis based algorithms are not able to discriminate between inter and intra-class variations. If we compare the performance of automatic algorithms with human performance (Fig. 2, humans are similar to the automatic algorithms. This analysis is not in agreement with the previous results by O'Toole *et. al.* [9]. Note that the previous study suggested that face recognition algorithms surpass humans matching faces over changes in illumination whereas, this study focuses on look-alike faces. We analyzed all the response from algorithms as well as humans and observed that in general, algorithms and humans provide similar results. However, for easier cases, human were able to give better results.

3.2. Texture Descriptor based Automatic Face Recognition Evaluation

Three texture descriptor based algorithms are also used for performance comparison, namely: Local Binary Pattern (LBP), Extended Uniform Circular LBP (EUCLBP) and Speeded Up Robust Feature (SURF) descriptors. LBP [10] encodes the texture of an image and uses χ^2 distance measure to compute match scores. Among several improvements over LBP, EUCLBP [11] has shown significant improvement. SURF [12], a faster version of Scale Invari-

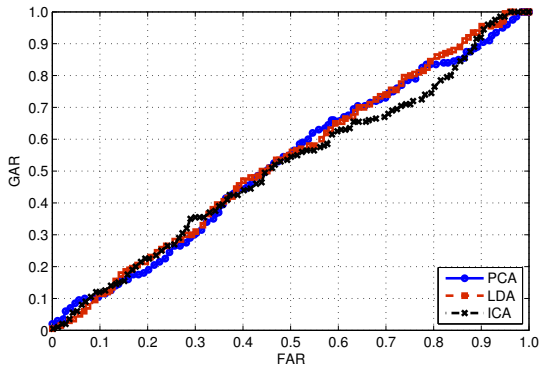


Figure 4. ROC plots for PCA LDA and ICA on the look-alike face database.

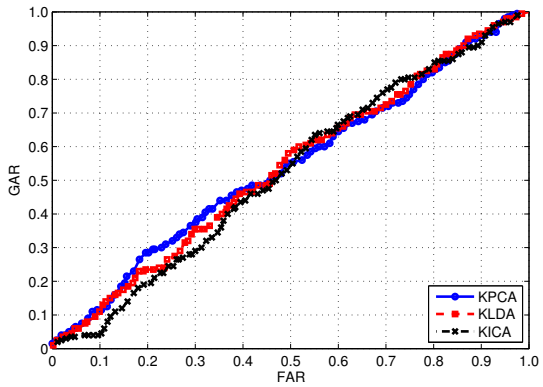


Figure 5. ROC plots for KPCA, KLDA, and KICA on the look-alike face database.

ant Feature Transform (SIFT) [13], is also used as an effective approach for face recognition. With same experimental setup, these three algorithms are trained and then used for performance evaluation on look-alike database. During training-testing (on the combined database), we observe that these three algorithms are at least 2% better than any subspace based algorithms.

Similar to subspace based algorithms, ROC curves for texture descriptor based algorithms are computed for look-alike face database. In this experiments, we observe that, in terms of verification performance, there is no significant difference between texture based algorithms and subspace based algorithms as well as human performance. This observation suggests that though existing algorithms may yield good accuracy on pose, expression, and illumination variations, challenging covariates such as *look-alikes* poses a major challenge.

4. Summary

For face recognition (even for object recognition), feature extractor should minimize the intra-class differences and maximize the inter-class variations. However, the presence of covariates such as look-alikes significantly increase the intra-class variation. Performing recognition with such images is a challenge faced by both humans and automatic face recognition algorithms. This research explores the impact of an important but unexplored challenge, namely look-alikes, on the performance of human and automatic face recognition. We have prepared a look-alike face database and analyzed the human performance with the help of 50 volunteers. Further, for automatic algorithms, both subspace (or appearance) and texture descriptor based algorithms are used. The results suggest that, for look-alikes, humans and automatic algorithms does not perform better than simple coin tossing (almost 50% probability of correct classification). We believe that these results may motivate the researchers to start considering complex covariates including *look-alikes*.

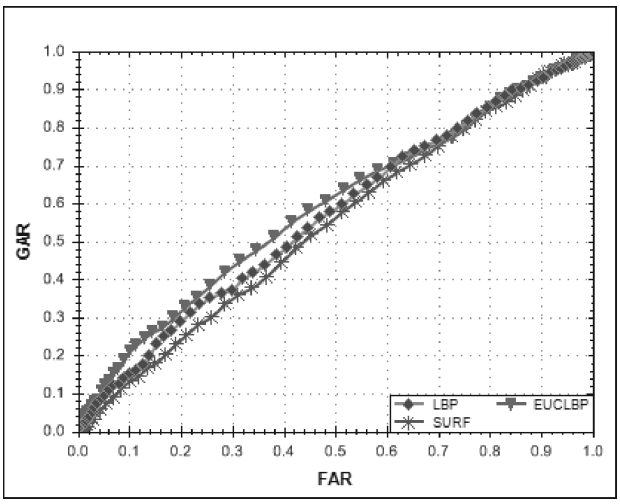


Figure 6. ROC plots for texture descriptor based approaches (LBP, EUCLBP, and SURF) on the look-alike face database.

5. Acknowledgment

The authors would like to thank the volunteers to participate in this research.

References

[1] M. Kosmerlj, T. Fladsrud, E. Hjelms, and E. Snekkenes, "Face recognition issues in a border control environment," in *Advances in Biometrics - Lecture Notes in Computer Science*, vol. 3832, pp. 33–39. Springer, 2005. 1

[2] S. Carpenter, "Why do 'they all look alike'?," 2000, Monitor on Psychology. 1

[3] N. Furl, P. Jonathon Phillips, and A.J. O'Toole, "Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis," *Cognitive Science*, vol. 26, no. 6, pp. 797 – 815, 2002. 1

[4] B. Schölkopf, A.J. Smola, and K.-R. Müller, *Kernel principal component analysis*, pp. 327–352, MIT Press, Cambridge, MA, USA, 1999. 4

[5] A.M. Martinez and A.C. Kak, "Pca versus lda," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 2, pp. 228 –233, Feb. 2001. 4

[6] B.A. Draper, K. Baek, M.S. Bartlett, and J.R. Beveridge, "Recognizing faces with pca and ica," *Computer Vision and Image Understanding*, vol. 91, pp. 115–137, 2003. 4

[7] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615–1618, 2003. 4

[8] P. J. Flynn, K. W. Bowyer, and P. J. Phillips, "Assessment of time dependency in face recognition: an initial study," in *Proceedings of Audio- and Video-Based Biometric Person Authentication*, 2003, pp. 44–51. 5

[9] A.J. O'Toole, P.J. Phillips, Fang Jiang, J. Ayyad, N. Penard, and H. Abdi, "Face recognition algorithms surpass humans matching faces over changes in illumination," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 9, pp. 1642 –1646, sep. 2007. 5

[10] T. Ahonen, A. Hadid, and M. Pietikinen, "Face description with local binary patterns: application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006. 5

[11] H.S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "On matching sketches with digital face images," in *IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*, 2010, pp. 1 –7. 5

[12] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, pp. 346–359, 2008. 5

[13] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal on Computer Vision*, vol. 60, pp. 91–110, 2004. 6