Face Recognition for Look-Alikes: A Preliminary Study

Hemank Lamba, Ankit Sarkar, Mayank Vatsa and Richa Singh
IIIT Delhi, India
{hemank08025, ankit08009, mayank, rsingh}@iiitd.ac.in

Afzel Noore
West Virginia Univeristy, USA
afzel.noore@mail.wvu.edu

Abstract

One of the major challenges of face recognition is to design a feature extractor and matcher 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 subject 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-alike faces and their effect on human performance and automatic face recognition algorithms. There is three fold contribution in this research: firstly, we analyze the human recognition capabilities for look-alike appearances. Secondly, we compare human recognition performance with ten existing face recognition algorithms, and finally, proposed an algorithm to improve the face verification accuracy. The analysis shows that neither humans nor automatic face recognition algorithms are efficient in recognizing look-alikes.

1. INTRODUCTION

Humans effortlessly process information obtained from multiple sensory inputs and have the ability to recognize individuals even with limited correlation, redundant information, or when certain features appear partially hidden, camouflaged or disguised [17]. 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 [14]. 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.

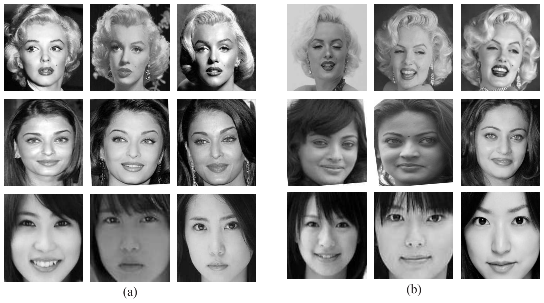

Figure 1. Examples of look-alikes: (a) genuine and (b) look-alikes of respective individuals in (a).

As shown in Figure 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 subspace, point/feature, and may be, texture. This assertion is based on the study by Kosmerlj *et. al.* [8]. In this study, an experiment was conducted to estimate the percentage of Norwegian people having one or more look-alikes. The 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 races look alike. Carpenter [4] 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 a person whether he is African, Asian or Hispanic and they do not explore the 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 proposed by Furl *et al.* [7] suggests that other race effect occurs as a result of greater experience we have with own- versus other-race faces. In another research by Phillips *et al*. [12], it is suggested that face recognition algorithms find it difficult to differentiate among twins. Twins can be also be considered as biological look-alikes and therefore, in our opinion, it is equally challenging.

Many law enforcement applications have to deal with this important challenge. The challenge of look-alikes is studied by the cognitive scientists but no proper evaluation has been performed for automatic algorithms. Contributions of this paper are therefore (1) analyzing human recognition capabilities for look-alike appearances, and (2) comparing it with several existing face recognition algorithms. For automatic face recognition, 10 different algorithms are compared including kernel subspace approaches, their linear counterparts, and texture descriptors. Further, a face recognition algorithm is proposed that improves the verification performance significantly compared to existing algorithms.

2. Human Recognition Capability for Look-alike Faces

To the best of our knowledge, there is no study that evaluates human capabilities as well as automatic 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 a database. However, different websites present 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 origins) and their look-alikes[1]. Each subject/class has 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 should 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].

2.2. Human Evaluation Protocol

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

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

The volunteers were shown face images from the *look-alike database* and they had to recognize and find genuine pairs. Similar to [11], they were shown easy as well as difficult pairs obtained using Principal Component Analysis (PCA)[3]. For every pair, the volunteers were asked to submit the response as (1) they are same or (2) they are not same.

Besides this, for every pair, volunteers were given a specific time of 20 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 20 seconds to look at the individual's face and document. To better analyze the results, we also asked the volunteers to mention if they know the given pair from past or not (familiar vs. unfamiliar faces). Finally, volunteers were asked to mention what specific features they used for recognizing faces.

2.3. Results and Analysis of Human Evaluation

On average, 56.6% human responses were found to be correct. With familiar faces, the average accuracy is 66.4% whereas for unfamiliar faces, the average accuracy is 53.5%. Key results are summarized below:

- The responses suggest that human verification accuracy for the 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. In our evaluation, interestingly, humans performed better on male classes compared to female classes.
- For some easy cases, volunteers easily performed correct verification whereas for complex cases, most volunteers were not able to correctly verify. As shown in

[1] Some look-alikes are genuine and some are intentional. Since the images are downloaded from various sources, it is difficult to classify them into these two categories. Therefore, irrespective of genuine or intentional, we use these images to study look-alike recognition.

[2] Download instructions are available at http://research.iiitd.edu.in/groups/iab/resources.html

[3] PCA score were used to rank the level of difficulty for all the image pairs. The pair with the highest score was considered as the easiest pair and the one with the lowest score was considered as the most difficult pair.

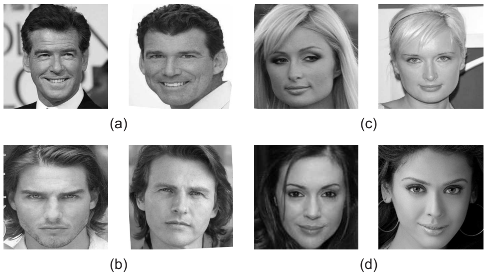

Figure 2. Examples of image pairs - images of two different individuals (look-alikes) - that are shown to the volunteers. These images are ordered in increasing order of complexity, i.e. the first pair is easy and the last pair is difficult.

Figure 2, the first image pair was considered as easy and the last pair was the most complex where volunteers made mistakes.

- In the experiments, *other race effect* did not affect the human performance. However, it was observed that images of western origin are easier to match compared to other races.
- Since the volunteers were requested to mention about the (un)familiarity with the image pairs, it was possible to analyze the effect of familiar vs. unfamiliar face recognition in humans. Generally, it is assumed that humans are very good at familiar face recognition. However, t-test at 90% confidence interval shows that there is no significant difference between unfamiliar and familiar face verification performance with look alike variations. This is an interesting result and, to the best of our knowledge, is not reported elsewhere.
- It was also noted that out of the 979 responses in human evaluation, timeout occurred only 37 times. A timeout means that the person was unable to judge the similarity within the allotted 20 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 an important role in making a decision. However, few responses also suggested that overall face appearance was important.

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 the look-alike database.

3.1. Automatic Face Recognition Algorithms

Among various techniques, subspace based face verification approaches have received major attention. These algorithms generally use subspace analysis methods to address pose, expression, and illumination variations. Examples of these algorithms include Principal Component Analysis, Linear Discriminant Analysis (LDA), and Independent Component Analysis (ICA). According to the Vapnik-Chervonenkis theory, mappings from lower dimensional space (input space) to higher dimensional space, in general, provides increased classification capabilities [13]. However, increasing the dimensions can increase the computational complexity. Kernel tricks can be used to overcome this issue and still get the benefits of higher dimensional mapping. In face recognition, past research has shown that manifolds can be discriminating but with kernels, 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).

Texture descriptor based algorithms are also used for performance comparison, namely: Local Binary Pattern (LBP), Extended Uniform Circular LBP (EUCLBP), Speeded Up Robust Feature (SURF) descriptors, and dynamic feed-forward neural network architecture based 2D log polar Gabor transform (GNN) [16]. LBP [1] encodes the texture of an image and uses χ^2 distance to compute the match scores. Among several improvements over LBP, EUCLBP [3] has shown significant improvement. SURF [2], a faster version of Scale Invariant Feature Transform (SIFT) [9], is also used as an effective approach for face recognition. Finally, GNN is a relatively new algorithm which encode binary texture pattern and matching is performed using Hamming distance measure. The authors have shown that the GNN algorithm outperforms existing algorithms on disguise cases.

3.2. Experimental Protocol and Results

Before evaluating on the look-alike database, the algorithms are first trained and evaluated on publicly available databases. The experiments on PCA-KPCA, ICA-KICA, LDA-KLDA, LBP, EUCLBP, SURF, and GNN 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. Table 1 lists the databases used and the number of classes selected from the individual databases. The combined database contains over 10,000 images pertaining to 600 subjects. The database is divided into two sets: (1) training dataset and (2) gallery-probe dataset. The train-

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

Face Database	Number of Classes (subjects)
AR [10]	120
CMU - PIE [15]	65
Notre Dame [6]	315
Equinox [5]	100
Total	600

Table 2. Verification accuracy of face verification algorithms on combined face database (at 0.1% FAR).

Algorithm	Verification Accuracy (%)
PCA	61.4
KPCA	77.2
ICA	62.7
KICA	71.6
LDA	73.0
KLDA	78.8
LBP	80.9
EUCLBP	82.1
SURF	82.7
GNN	83.1

ing dataset is used to train the individual algorithms and it comprises images pertaining to 40% subjects (i.e. over 4,000 images). The gallery-probe test dataset contains the remaining 60% subjects (i.e. over 6,000 images) and is used for performance evaluation. Note that, in the experiments, the training and testing datasets are not overlapping. This means that all the individuals in the testing data are unseen. The train-test partitioning is repeated 10 times and verification accuracies are reported at 0.1% FAR. The optimal parameters for algorithms, for example, the kernel parameters in non-linear subspace learning approaches, are obtained empirically.

As shown in Table 2, it is observed on the combined database that non-linear kernel algorithms can better encode the facial features compared to their linear counterparts and KLDA outperforms other subspace-based approaches. Further, the texture based algorithms provide are at least 2% higher verification accuracy than subspace based algorithms.

After training-testing on the non-homogeneous combined face database, trained algorithms are evaluated using the look-alike face database. Figures 3, 4, 5 and 7 show the Receiver Operating Characteristics (ROC) curves 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 in the range of 45-50% which is not better than simple coin tossing. This is mainly because with look-alike face images, the algorithms are not able to discriminate between the inter and intra-class variations. Further, in terms of verification performance, there is no significant difference between texture and subspace-based algorithms. On comparing with human responses, it is observed that most of the response from algorithms as well as humans are, in general, similar. However, for easier cases, humans provide better results. This observation suggests that though existing algorithms may yield good accuracy on pose, expression, and illumination variations, challenging covariates such as *look-alikes* pose a major challenge.

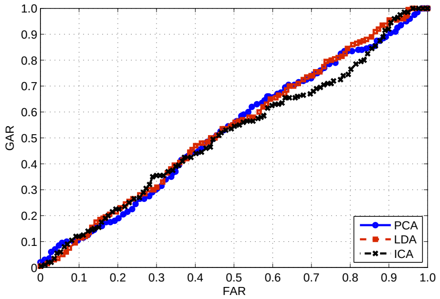

Figure 3. ROC plots for PCA LDA and ICA on the look-alike face database.

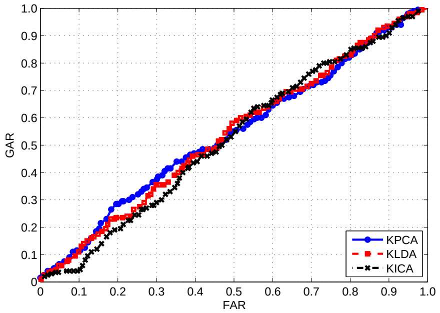

Figure 4. ROC plots for KPCA, KLDA, and KICA on the look-alike face database.

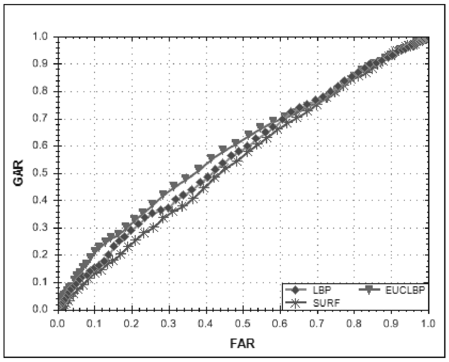

Figure 5. ROC plots for texture descriptor based approaches (LBP, EUCLBP, and SURF) on the look-alike face database.

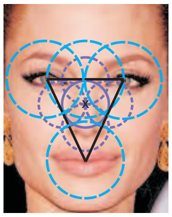

Figure 6. Facial regions selected using eyes-mouth coordinates.

4. Proposed Approach for Verification of Look-alike Faces

The motivation behind the proposed approach is based on the concept that human mind performs recognition at different levels of abstraction and use local and global facial regions [17]. In the proposed approach, features are extracted from overlapping facial regions and the resulting information is combined at the match score level to perform verification. With this approach, different facial regions such as nose, ears, or combination of two or more features are analyzed and used for matching. The proposed algorithm is described as follows:

- *Facial Regions from Face Image*: Face detection using Adaboost face detection is used to detect face and eyes-mouth coordinates. The detected face images serve as the global face region for feature extraction and eye-mouth coordinates are used for region selection. From human analysis, it has been observed that eyes and mouth regions are the most discriminating. As shown in Figure 6, to incorporate this result in the automatic algorithm, local facial regions are extracted from the eyes and mouth region. The innermost circular region is selected based on the triangle connecting eyes-mouth coordinates. Using the same radius, surrounding local regions are selected (one in each quadrant in the innermost circle). These circular regions are selected to accommodate eyes-mouth region. As shown in Figure 6, four regions are selected using the inter-eye distance as the diameter: (1) inter-eye region, (2) left eye region, (3) right eye region, and (4) mouth region. With this approach, the global and nine local facial regions are obtained.

- *Feature Extraction*: The next step in the proposed algorithm is feature extraction from face regions. Here, we use GNN for extracting binary phase features [16] from each facial region. The feature extractor extracts binary phase features from each face region and Hamming distance measure is used to calculate the match score of two binary phase features. For a given face image, the feature extraction process is performed separately for all 10 face regions, i.e. computation of 10 phase features, one for each region. In other words, the algorithm is used to generate gallery phase features corresponding to the gallery face images and are stored in the database. During query, phase features for the probe image is generated and matched using Hamming distance. The algorithm finally generates match score vector, $\mathbf{s}$, where each element s_i, $(i = 1, \cdots, 10)$ is associated with one of the 10 face regions. Each of these match scores are in the range [0, 1] where 0 represents perfect accept and 1 represents perfect reject.

- *Classification*: The final step in the proposed face verification algorithm is classification of match score vector $\mathbf{s}$ to yield an output decision of *genuine* or *impostor*. For look-alike cases, different facial regions may provide conflicting decisions. For example, full face region may reject a genuine subject but some of the local regions may provide a decision to accept. In such cases, a non-linear classification algorithm is required that can efficiently address these confounding match scores. In this research, Support Vector Machine (SVM) [18] is used as a two-class classification algorithm. The match score vector $\mathbf{s}$ is used as input to the SVM classifier. Since SVM requires training, the match scores and their labels obtained from the training database are used to train SVM for classification. The optimal hyperplane which separates the complete training data into two different classes in the higher dimensional feature space can be obtained using SVM learning [18].

  In the testing phase, the match score vector obtained by matching the gallery and probe pair, $\mathbf{s}_{probe}$ is classified using the trained SVM. To verify the identity, a decision to *accept* or *reject* is made on the test pattern using a threshold t,

$$Decision(s_{probe}) = \begin{cases} Accept, & \text{if SVM output} > t \\ Reject, & \text{otherwise.} \end{cases} \quad (1)$$

4.1. Results and Analysis

Similar to previous experiments, same experimental setup is used to train the proposed algorithm and then used for performance evaluation on the look-alike database. For SVM, best results are obtained using the radial basis function with kernel parameter = 4. On the combined database, the proposed algorithm yields an accuracy of 86.7% which is at least 3.6% better than existing algorithms.

ROCs in Figure 7 show that the proposed algorithm yields significantly improved performance on the look-alike database. The proposed algorithm yields an EER of 41% which is about 5-10% better than other algorithms. In terms of verification accuracy at 0.1% FAR, the proposed algorithm improves the performance at least two times compared to GNN and other existing algorithms. One can argue

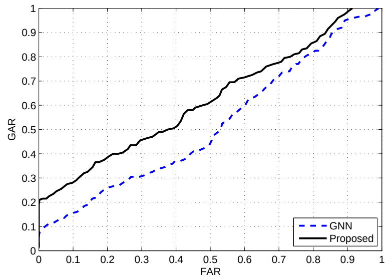

Figure 7. ROC plots for GNN and the proposed algorithm on the look-alike face database.

that any feature extractor can be applied in the proposed algorithm. To substantiate this, experiments were performed with all other feature extractors (used in Section 3.1) by applying zero-padding on global and local regions and an improvement of 2-5% was observed. This substantiates the motivation/hypothesis that both local and global facial regions are important in face recognition.

5. Summary

For face 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 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 do not perform better than random guess. We also proposed an algorithm that significantly improves the performance compared to existing algorithms. However, we believe that it is important to start considering complex covariates including *look-alikes* and develop advanced algorithms.

6. Acknowledgment

The authors would like to thank the volunteers to participate in this research. This research is supported through a grant from Department of Information Technology, India.

References

[1] T. Ahonen, A. Hadid, and M. Pietikinen. Face description with local binary patterns: application to face recognition. *IEEE Transactions on PAMI*, 28(12):2037–2041, 2006. 3

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110:346–359, 2008. 3

[3] H. 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*, pages 1 –7, 2010. 3

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

[5] http://www.equinoxsensors.com/products/HID.html. 4

[6] 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*, pages 44–51, 2003. 4

[7] N. Furl, P. J. Phillips, and A. O'Toole. Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. *Cognitive Science*, 26(6):797 – 815, 2002. 2

[8] 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*, volume 3832, pages 33–39. Springer, 2005. 1

[9] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60:91–110, 2004. 3

[10] A. Martinez and R. Benavente. The AR face database, 1998. Computer Vision Center, Technical Report. 4

[11] A. O'Toole, P. Phillips, F. Jiang, J. Ayyad, N. Penard, and H. Abdi. Face recognition algorithms surpass humans matching faces over changes in illumination. *IEEE Transactions on PAMI*, 29(9):1642 –1646, sep. 2007. 2

[12] P. Phillips, P. Flynn, K. Bowyer, R. Bruegge, P. Grother, G. Quinn, and M. Pruitt. Distinguishing identical twins by face recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 185 –192, 2011. 2

[13] B. Schölkopf, A. Smola, and K.-R. Müller. *Kernel principal component analysis*, pages 327–352. MIT Press, Cambridge, MA, USA, 1999. 3

[14] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on PAMI*, 29:411–426, 2007. 1

[15] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE Transactions on PAMI*, 25(12):1615–1618, 2003. 4

[16] R. Singh, M. Vatsa, and A. Noore. Face recognition with disguise and single gallery images. *Image and Vision Computing*, 27(3):245–257, 2009. 3, 5

[17] P. Sinha, B. J. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: 19 results all computer vision researchers should know about. *Proceedings of IEEE*, 94(11):1948–1962, 2006. 1, 5

[18] V. Vapnik, S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation and signal processing. *Advances in Neural Information Processing Systems*, 9:281–287, 1997. 5