A. Brömme, C. Busch, N. Damer, A. Dantcheva, M. Gomez-Barrero, K. Raja,
C. Rathgeb, A. Sequeira and A. Uhl (Eds.): BIOSIG 2021,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021

Identical Twins as a Facial Similarity Benchmark for Human Facial Recognition

John McCauley[1], Sobhan Soleymani[1], Brady Williams[1], John Dando[1],
Nasser Nasrabadi[2], Jeremy Dawson[2]

Abstract: The problem of distinguishing identical twins and non-twin look-alikes in automated facial recognition (FR) applications has become increasingly important with the widespread adoption of facial biometrics. This work presents an application of one of the largest twin datasets compiled to date to address two FR challenges: 1) determining a baseline measure of facial similarity between identical twins and 2) applying this similarity measure to determine the impact of doppelgangers, or look-alikes, on FR performance for large face datasets. The facial similarity measure is determined via a deep convolutional neural network. The proposed network provides a quantitative similarity score for any two given faces and has been applied to large-scale face datasets to identify similar face pairs.

Keywords: Facial Similarity, Facial Recognition, Identical Twins, Look-alikes.

1 Introduction

Identical, or monozygotic, twins continue to pose significant challenges to facial recognition (FR) systems. Paone et al. [Pa14] found that the average face recognition system has a significantly higher equal error rate (EER) when presented with a population of identical twins versus a non-twin population. As any facial dataset becomes large, the probability of look-alikes becomes larger, giving a higher likelihood of false matches. However, for identical twins and look-alikes, a high comparison score may not be directly correlated to human-perceived facial similarity due to transformations applied and features extracted by the FR algorithm.

The primary purpose of this work is to develop and apply a similarity measure to evaluate baseline facial similarity between identical twins as a worst-case scenario of similarity in face comparison. This work is motivated by two factors 1) the need to better understand the relationship between facial similarity and the comparison score returned by a FR system (i.e., the difference between FR and facial similarity) and 2) the need for isolating potential look-alikes in large face datasets to estimate the frequency of look-alike occurrence in any dataset.

[1] West Virginia University, Lane Department of Computer Science and Electrical Engineering, Morgantown, West Virginia, USA, {jamccauley, ssoleyma, bwwilliams, jmdando}@mix.wvu.edu
[2] West Virginia University, Lane Department of Computer Science and Electrical Engineering, Morgantown, West Virginia, USA, {nasser.nasrabadi, jeremy.dawson}@mail.wvu.edu

This work will present: 1) an analysis of the performance of one commercial off the shelf (COTS) FR tool and one academic FR algorithm on one of the largest identical twin databases to date, as well as two large non-twin datasets, 2) a convolutional neural network (CNN) based measure for quantifying facial similarity that can inform how facial similarity is related to comparison score, and 3) an application of this similarity measure to determine the baseline similarity of identical twin pairs and identify non-twin look-alikes in a large face dataset.

2 Background

In previous studies of twin recognition, the challenge of distinguishing between two identical twins has been explored for face, fingerprint, and iris modalities [Ne12], [RM13], [BF16], [Su10]. While face recognition is one of the most widely used biometric modalities, it also faces the biggest challenge when presented with identical twins. Two of the earliest studies on the face recognition of twins [Ph11] & [Pr11] found that several COTS face matchers could identify twins when imaging conditions were ideal (i.e., studio lighting, neutral expression), but performance was measurably decreased as imaging conditions were varied [Ph11]. The goal of the work presented herein is not intended to explore the performance of FR systems on identical twins. Instead, this work investigates the related topic of facial similarity using identical twin pairs as a worst-case baseline of facial similarity. Previous studies of facial similarity have drawn an important distinction between face recognition and facial similarity. An early study identified the topic of facial similarity to be distinct from that of face recognition and developed a facial similarity measure based on an Eigenface framework [RCC04]. A recent work from Sadovnik et al. [Sa18] explored a CNN approach to rank similar faces within a dataset, showing evidence that facial similarity is highly related to, but distinct from, facial recognition. This work used hand selected similar face pairs to train a neural network to accomplish the distinct task of facial similarity determination.

Another application of facial similarity determination is the selection of faces for morph generation. Röttcher et al. [RSB20] present a method of determining facial similarity using a variety of factors, showing that their intelligent morph pair selection produces better morphs than random selection.

Accurate identification of look-alikes poses a similar challenge to that of identical twin recognition. Work from Kosmerlj et al. [Ko05] on the effect of look-alikes on a border control FR application found that the system evaluated would not be robust to the occurrence of look-alikes based on their estimated frequency of look-alike individuals. Studies presented in [La11] & [STN15] found FR performance on look-alikes to be very low for feature-based and deep learning algorithms. These works motivate need for the ability to identify potential look-alikes in any given dataset to better evaluate the hardest cases presented to FR systems. Here, we seek to expand upon previous research in this sphere by determining a baseline facial similarity measure for any two faces based on the similarity of identical twins.

2.1 Datasets & Match Performance Analysis

The face image data used in this work comes from multiple sources. The first of these is a twin dataset[1] that contains 2,269 unique identities, 1,438 of which are identical twins. The remaining portion of the dataset is comprised of fraternal twins, relatives to the twin pairs, and non-twin participants. The second dataset is comprised of face images of participants from the general public (i.e., non-twins) with 5,295 unique identities. A third dataset was constructed using the second, non-twin dataset combined with the CelebA dataset [Li15], resulting in a dataset that contains a total of 15,455 unique identities.

The initial task of the work presented here was to analyze the performance of two facial recognition systems on the datasets described above. Two FR tools were used in this experiment, the first of which was a COTS "black box" matcher, and the second was the FaceNet matcher [SKP15], which is based on the Inception-ResNet v1 architecture. The first experiment was a baseline analysis of the matchers when presented with only the identical twin pairs from the twin dataset to determine the effect of highly similar faces on the non-mated distribution of a FR experiment. In addition, a mated comparison was made for each identity to show the relationship between the identical twin non-mated distribution and mated distribution. The mean comparison score of the identical twins in this baseline experiment is used as the threshold for each of the remaining comparison experiments. This score represents an experimental threshold for individuals with high facial similarity, and is used later in this paper to extract potential look-alikes in our dataset for further analysis by the proposed similarity network. For the remaining comparison experiments, the approach presented in Howard et al. [HSV19] was used, wherein all-to-all matching was performed on each of the face datasets retaining only the non-mated, or impostor, distributions. The remaining comparison experiments correspond to each of the face datasets used in this work and are as follows: twin dataset, non-twin dataset, and the combined non-twin and CelebA dataset.

2.2 Similarity Network

To determine a quantitative measure of similarity between identical twins, a deep CNN was implemented. This network was designed with a twin architecture (also known as Siamese architecture) to directly compare two faces. Each half of this network was comprised of a FaceNet architecture, with the weights of the network shared between the two halves, as shown in Figure 1.

[1] This is the first use of the twin dataset; it is available upon request.

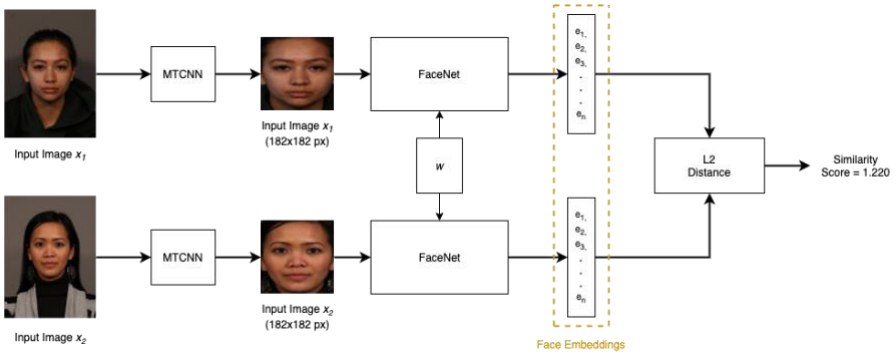

Fig. 1: Similarity network diagram.

The FaceNet architecture was chosen as the foundation of this network for its high accuracy on typically difficult, similar face images, as well as its ability to generate highly representative face embeddings. In [SKP15], the network is trained to generate embeddings which directly correspond to facial similarity in the embedded feature space. This is advantageous, as our work seeks to quantify the similarity of identical twin pairs rather than simply generating a comparison score. The network was optimized using the contrastive loss function [HSL06] to minimize the L2 distance between similar samples who reside close to one another in the feature space while maximizing the L2 distance between the dissimilar samples. The output of this network consists of the L2 distance between two samples in the feature space. As this calculation gives similar samples a low score and dissimilar samples a high score, for clarity, the scores are inverted such that similar samples have a high score and dissimilar samples a low score. This was achieved by subtracting each resultant similarity score from the maximum similarity score in a given set of scores. The training phase of the network consisted of fine tuning the weights with data from the twin dataset. Starting with the network pre-trained on the VGGFace2 dataset, the network was fine-tuned on a subset of the twin database. This fine tuning was performed on a tailored verification task where a pair of identical twin images represents the positive case, and a pair of unrelated look-alikes represents the negative case. This training encouraged the network to group together those samples with the facial similarity of identical twins in the embedded feature space, and inversely pushed apart those samples not as similar as identical twins.

The training and testing dataset for this network was comprised of a subset of the twin dataset. This dataset contained images of identical twin pairs and non-mated look-alikes to the twin pairs sorted into an equal number of mated and non-mated pairs, where a mated pair consists of (Twin A vs. Twin B), and a non-mated pair consists of (Twin A vs. look-alike). The look-alikes for each identity were found by selecting the identities with the highest FaceNet comparison score to each twin identity. This training schema was chosen because the network should learn to determine facial similarity from the most similar face pairs available (i.e., identical twins). It is expected that an individual's identical twin will be more similar than any potential look-alike, as such, the network is trained to identify the face pairs with the highest facial similarity. The dataset contains 645 identical twin

identities, with a total of 3,203 images, split 80/20 for training and testing.

3 Results

3.1 Match Performance Analysis

Figures 2 illustrates the results of the identical twin baseline experiments, indicating that the average comparison score for identical twins trends higher than the comparison score for non-twin matches. The mean comparison score for identical twins in this baseline represents the experimental twin threshold *T*, and, when compared to the mated score distribution, this threshold approximates the left tail of the mated distribution for both matchers tested. All-to-all impostor or non-mated matching was performed using both matchers. The comparison scores for each of these experiments were analyzed to extract the scores falling at and above the experimental twin threshold, *T*, presented in Table 1. In each of the matching experiments, it is shown that an overwhelming majority of comparison scores fall below the experimental twin threshold, indicating that non-mated look-alikes are a rare occurrence in the population used in this study. Due to the relatively small number of identities in the datasets used for this evaluation, it is not possible to accurately predict the frequency of look-alike occurrence in general from these results. However, the similarity measure described in the next section provides a method of finding highly similar faces in any given dataset.

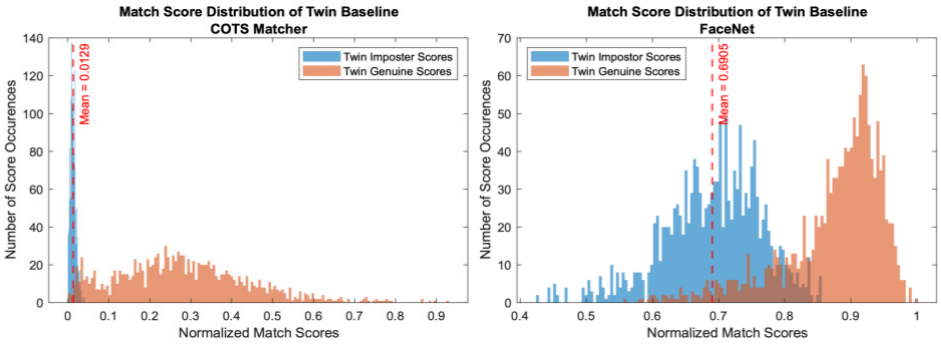

Fig. 2: Twin baseline match experiment results from COTS (left) and FaceNet (right) matchers. The red line shows the mean comparison score of identical twins.

Dataset	Relationship	# Scores >= *T*	Avg. Score	Score Range	% of Matches
Twin Dataset – COTS Matcher	Ident. Twin	713	0.0188	[0.0129-0.0431]	0.0139%
	Family Member	50	0.0197	[0.0135-0.0344]	0.00097%
	No Relation	199	0.0137	[0.0129-0.0189]	0.0038%
Twin Dataset – FaceNet Matcher	Ident. Twin	868	0.746	[0.6905-0.856]	0.0168%
	Family Member	50	0.753	[0.6905-0.83]	0.00097%
	No Relation	6	0.702	[0.694-0.7177]	0.00012%
Non-twin dataset – COTS Matcher	No Relation	16274	0.0144	[0.0129-0.0283]	0.0580%
Non-twin dataset – FaceNet Matcher	No Relation	97	0.704	[0.6905-0.76]	0.000346%
Large Scale Non-twin dataset – FaceNet* Matcher	No Relation	792	0.71	[0.6905-0.76]	0.000331%
	* the large-scale non-twin dataset experiment was performed exclusively on the FaceNet matcher				

Tab. 1: Matching analysis experiments, comparison scores above the twin threshold.

3.2 Similarity Network

After training and testing, the proposed network achieved a train AUC of 0.917, and test AUC of 0.979 in the classification of a pair of face images as a twin pair or look-alike pair. While the end goal of this network is not verification, the accuracy of the network on the tailored verification task shows that the network can accurately identify similar face pairs. This similarity network was then applied to both the twin dataset and large-scale non-twin dataset to observe the general similarity of twin and non-twin individuals. Initially, the similarity score of only the identical twin pairs was calculated (Figure 3). This distribution of similarity scores for identical twin pairs is the foundation of the worst-case baseline measure of similarity. As identical twins exist on a spectrum of similarity, two measurements of the baseline similarity between identical twin pairs are reported. The mean similarity score between identical twin pairs, 1.09, captures the similarity of both highly similar and dissimilar twins, while the fourth quartile of the similarity score distribution, ≥1.29, represents only the most similar twin pairs. In this experiment, the fourth quartile score of the distribution may more accurately represent the worst case of similarity presented to FR systems.

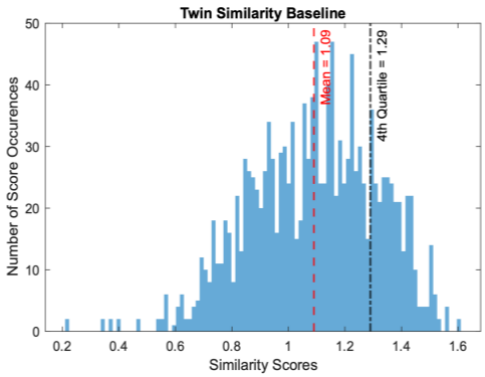

Fig. 3: Identical twin only similarity baseline. The red line shows the mean of the twin similarity distribution, and the black line the fourth quartile of the distribution.

After determining the worst-case baseline for facial similarity, this measure was used to set the threshold for the similarity scores of the large-scale non-twin dataset. Since the network was fine-tuned using only ideal face images, the similarity score returned for "in-the-wild" face images may not be as robust as the similarity score returned for controlled images. Several examples of identical twin pairs and non-mated pairs with similarity scores exceeding the baseline measurements are shown in Figure 4.

Fig. 4: Examples of identical twin pairs, non-twin look-alikes, and dissimilar face comparisons as determined by the similarity scores from the similarity network.

An additional analysis was performed to correlate the comparison score results of the COTS matcher to the similarity score obtained from our similarity network. Using the non-mated pairs whose comparison scores exceeded the experimental twin threshold in the matching experiments detailed above, a comparison was made to the similarity score calculated for the same pairs. Examples of individuals with high COTS comparison scores and the corresponding similarity score are highlighted below (see Figure 5).

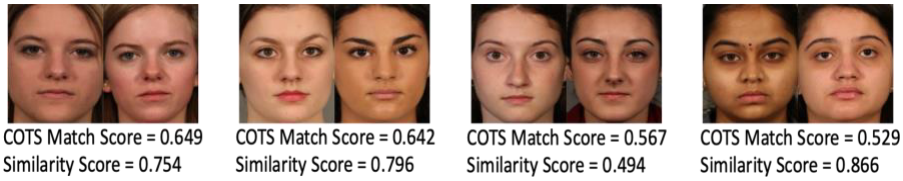

Fig. 5: High COTS comparison score, non-mated pairs and their corresponding similarity score.

As Figure 5 indicates, the COTS comparison score for each of these face pairs was high; however, none of the pairs' similarity scores are above the twin similarity threshold. This indicates that the comparison score returned by the COTS matcher is not directly correlated with facial similarity, and instead, may rely on other features of the image in its comparison score determination process.

Finally, an investigation into the number of potential look-alike pairs returned by the network while varying the similarity threshold was performed to further understand the occurrence of look-alikes in a given population of unrelated individuals (Fig. 6).

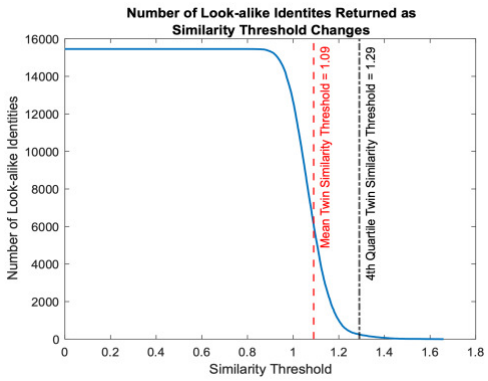

Fig. 6: Number of look-alike identities versus a range of similarity thresholds – large scale non-twin dataset.

Given the mean twin similarity threshold of 1.09, 6,153 of the total 15,455 identities in the large-scale non-twin dataset have at least one similarity comparison at or above the threshold. This means 39.8% of the identities have one or more potential look-alike at this level of similarity. At and above the fourth quartile threshold, only 228 identities have one or more potential look-alike, or 1.475% of identities in the dataset.

4 Conclusion

This work presents an application of one of the largest twin databases to date to better understand the challenges that look-alikes pose to biometric face recognition. Using this dataset, a baseline measure of the worst-case scenario of facial similarity in FR was

calculated using a deep CNN. Additionally a performance analysis of two FR tools presented with highly similar faces was carried out, to demonstrate the impact of highly similar faces on FR tools. Using an experimental twin threshold, potential look-alikes were extracted from the datasets for further analysis.

The similarity measure presented here has several applications in FR at large. First, this measure is one way to compare facial similarity to a comparison score from a FR system in order to better understand the impact that facial similarity has on FR. Second, this measure can be directly applied to large face datasets to identify potential look-alikes. Face pairs with high facial similarity can then be identified as difficult cases for a FR system or be used for other applications such as the selection of suitably similar faces for intelligent morph pair generation.

Future work in this area could further explore the relationship between the comparison score returned by a FR tool and the similarity score returned by the proposed similarity network. Another topic of interest in this sphere is a translation of the so called 'birthday paradox' to large facial datasets. Much like the birthday paradox seeks to calculate the probability of two people in a given population sharing a birthday, calculating the probability of two unrelated individuals having high facial similarity based on the number of identities in a dataset would be a useful measure as face datasets continue to grow in size. This measure could be used to estimate the number of look-alikes in the population at large, or determine the difficulty of large face datasets.

This material is based upon work supported by the Center for Identification Technology Research and the National Science Foundation under Grant No. 1650474.

References

[Pa14] Paone, J. R. et al.: Double Trouble: Differentiating Identical Twins by Face Recognition. IEEE Transactions on Information Forensics and Security vol. 9 no. 2, S. 285–295, 2014.

[Ne12] Nejati, H. et al.: Wonder ears: Identification of identical twins from ear images. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). S. 1201–1204, 2012.

[RM13] Ricanek, K.; Mahalingam, G.: Biometrically, How Identical Are Identical Twins?. Computer vol. 46 no. 3, S. 94–96, 2013.

[BF16] Bowyer, K. W.; Flynn, P. J.: Biometric identification of identical twins: A survey. In: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS). S. 1–8, 2016.

[Su10] Sun, Z. et al.: A study of multibiometric traits of identical twins. Biometric Technology for Human Identification VII vol. 7667, p. 76670T, 2010.

[Ph11] Phillips, P. J. et al.: Distinguishing identical twins by face recognition. In: 2011 IEEE International Conference on Automatic Face Gesture Recognition (FG). S. 185–192, 2011.

[Pr11] Pruitt, M. T. et al.: Facial recognition of identical twins. In: 2011 International Joint

Conference on Biometrics (IJCB). S. 1–8, 2011.

[RCC04] Ramanathan, N.; Chellappa, R.; Chowdhury, A. K. R.: Facial similarity across age, disguise, illumination and pose. In: 2004 International Conference on Image Processing (ICIP). vol. 3, S. 1999-2002, 2004.

[Sa18] Sadovnik, A. et al.: Finding your Lookalike: Measuring Face Similarity Rather than Face Identity. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). S. 2408–24088, 2018.

[RSB20] Röttcher, A.; Scherhag, U.; Busch, C.: Finding the Suitable Doppelgänger for a Face Morphing Attack. In: 2020 IEEE International Joint Conference on Biometrics (IJCB). S. 1–7, 2020.

[Ko05] Kosmerlj, M. et al.: Face recognition issues in a border control environment. Advances in Biometrics - Lecture Notes in Computer Science vol. 3832, S. 33-39, 2005.

[La11] Lamba, H. et al.: Face recognition for look-alikes: A preliminary study. In: 2011 International Joint Conference on Biometrics (IJCB). S. 1–6, 2011.

[STN15] Sun, X.; Torfi, A.; and Nasrabadi, N.: Deep Siamese Convolutional Neural Networks for Identical Twins and look-alike Identification. in Deep Learning in Biometrics, 2015.

[Li15] Liu, Z. et al.: Deep Learning Face Attributes in the Wild. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 3730–3738, 2015.

[SKP15] Schroff, F.; Kalenichenko, D.; Philbin, J.: FaceNet: A Unified Embedding for Face Recognition and Clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). S. 815–823, 2015.

[HSV19] Howard, J. J. ; Sirotin, Y. B.; Vemury, A. R.: The Effect of Broad and Specific Demographic Homogeneity on the Impostor Distributions and False Match Rates in Face Recognition Algorithm Performance. 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS). S. 1-8, 2019.

[HSL06] Hadsell, R.; Chopra, S.; LeCun, Y.: Dimensionality Reduction by Learning an Invariant Mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06). S. 1735–1742, 2006.