2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Lookalike Disambiguation: Improving Face Identification Performance at Top Ranks

Thomas Swearingen and Arun Ross
Computer Science and Engineering Department
Michigan State University
East Lansing, MI 48824
Email: {swearin3, rossarun}@cse.msu.edu

***Abstract*—A face identification system compares an unknown input probe image to a gallery of labeled face images in order to determine the identity of the probe image. The result of identification is a ranked match list with the most similar gallery face image at the top (rank 1) and the least similar gallery face image at the bottom. In many systems, the top ranked gallery images may look very similar to the probe image as well as to each other and can sometimes result in the misidentification of the probe image. Such similar looking faces pertaining to different identities are referred to as lookalike faces. We hypothesize that a matcher specifically trained to disambiguate lookalike face images when combined with a regular face matcher will improve overall identification performance. This work proposes reranking the initial ranked match list using a disambiguator especially for lookalike face pairs. This work also evaluates schemes to select gallery images in the initial ranked match list that should be re-ranked. Experiments on the challenging TinyFace dataset shows that the proposed approach improves the closed-set identification accuracy of a state-of-the-art face matcher.**

I. Introduction

Biometrics is the science of recognizing individuals using biological or behavioral traits such as face, fingerprint, iris, or gait [1]. There are two primary scenarios for recognition, *verification* and *identification*. In verification, a biometric sample associated with a claimed identity is compared to a known biometric sample associated with that identity to render a match or no-match decision. In identification, a probe biometric sample is compared against a set of labeled biometric samples (called a gallery) to produce a ranked match list. The ranked match list contains an ordering of the gallery samples based on how similar they are to the probe sample (with the rank 1 sample being judged by the matcher[1] as the most similar to the probe). In closed-set identification, the probe image corresponds to one of the identities represented in the gallery. This is unlike open-set identification, where the identity corresponding to the probe image may or may not be represented in the gallery.

The way a ranked list is used in an identification system, varies across applications. In some applications, the top k ranks in the ranked match list are reviewed by a human investigator – therefore, the rank k cumulative identification accuracy is more important than just the rank 1 identification accuracy (“lights-on mode”). In other applications, there is no human review, so the rank 1 identification accuracy is of utmost importance (“lights-out mode”).

[1]We use the term “matcher” to denote the biometric (*e.g.*, face) recognition system.

Fig. 1. Examples of lookalike pairs as judged by an automated face matcher. Face images are from the TinyFace dataset [3].

In the ranked match list, the correct match may not occur at rank 1. For example, if the correct match occurs at rank 5 (so still close to the top rank), then it is possible that the matcher was “confused” by similar-looking faces from rank 1 through rank 4. These *lookalike* faces may be a special case in the context of face recognition and one not well-handled by a general-purpose face matcher [2]. Figure 1 shows examples of lookalike face image pairs, *i.e.*, imposter face image pairs, which an automated face matcher judged as being very similar. These lookalike identities can be viewed as doppelgängers with respect to that matcher.

In this work, we address the issue of lookalikes in face recognition. We propose the use of **two matchers**, the first, a *general-purpose face matcher*, and the second, a *lookalike disambiguator*. The general-purpose (GP) matcher is trained just like a normal face recognition system for identification. The lookalike disambiguator (LD) is trained specifically to distinguish between lookalikes. The GP matcher compares the probe face image to all gallery face images to obtain the initial ranked match list. The LD is then used to rerank a subset of the ranked match list (the subset of matches to rerank is selected

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412063

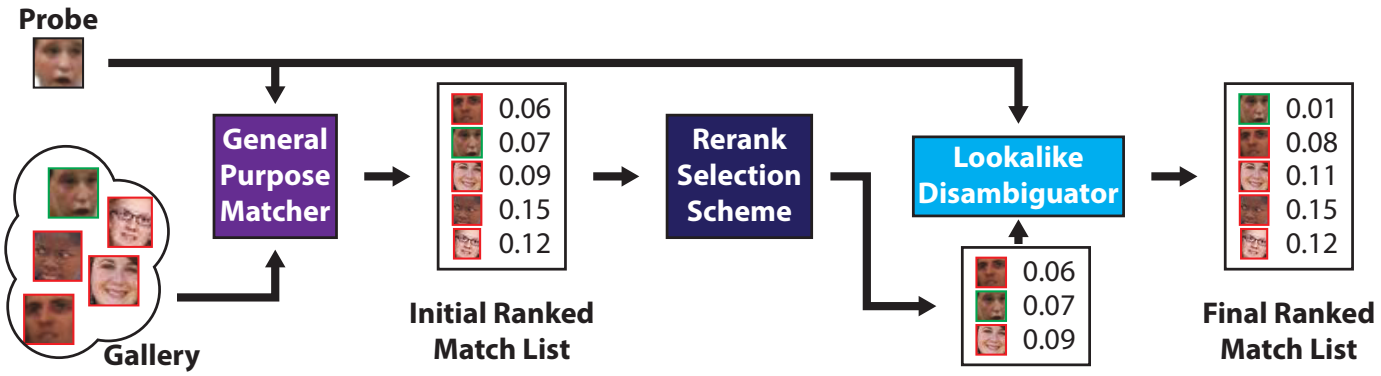


Fig. 2. Overview of the proposed approach. A general-purposed matcher compares a probe against all gallery samples to obtain an initial ranked match list (sorted in ascending order based on distance scores). A rerank selection scheme chooses a subset of gallery samples in the match list to rerank with the lookalike disambiguator. The lookalike disambiguator compares the probe with the selected gallery samples and outputs a new ranking. The final ranked match list contains the reranked gallery samples and the unselected gallery samples. The face images with a red border are of different subjects than the probe image, while the face images with a green border are of the same subject as the probe image.

adaptively based on the match scores). Figure 2 illustrates an overview of the proposed method.

This paper is organized as follows. Section II reviews related work. Section III explores the nature of matches in an identification search. Section IV describes the proposed reranking schemes and Section V the training approach for the GP matcher and the LD. Section VI describes the data used in this work and the lookalike discovery scheme. Section VII describes the experiments, Section VIII analyzes the results, and Section IX concludes the work.

II. Related Work

A. Lookalikes in Face Recognition

Analysis of automated face recognition systems reveals many underlying problems which increase the difficulty of successful matching. Some problems, like pose and illumination variations, are due to capture conditions [4] and have been well-explored in the literature [5]–[7]. Other problems, such as lookalikes, are inherent to the problem of face recognition itself for both humans and machines. The lookalike phenomenon has many causes, *viz.*, identical twins, kinship, surgical manipulations, doppelgängers, and others.

Identical twins were an early interest in the lookalike problem [8]–[11]. Identical twins, *i.e.*, monozygotic twins, have near-identical DNA. Although such twins have near-identical DNA and usually have very similar facial appearance, other factors such as environment and behavior may introduce differences in the facial appearance between them [12].

Since the facial appearance of identical twins is especially similar, some face recognition approaches for identical twins focus on specific aspects of the facial appearance, rather than the holistic approach adopted by many general-propose face matchers. For example, Srinivas *et al.* use facial marks (*e.g.*, freckles, scars, birthmarks) to tell identical twins apart [13]. One analysis showed that identical twins over age 40 are easier to differentiate by automated face recognition systems [11]. Le *et al.* propose face aging features to discern between identical twins [14]. However, not all approaches focus on a specific aspect of the facial appearance. For example, Sun *et al.* train a convolution neural network for distinguishing between identical twins [15].

Other research approach the lookalike problem more generally and do not focus strictly on distinguishing identical twins. Lambda *et al.* match face regions independently and fuse the results to facilitate lookalike disambiguation [16]. Smirnov *et al.* improve a general-purpose face matcher by maintaining a list of lookalikes to refine mini-batch selection [17]. Moeini *et al.* use 3D models to distinguish lookalike faces [18]. Sadovnik *et al.* point out that face *similarity* and face *identity* are related, but different concepts [2]. Classical face matchers are trained to distinguish between images from different identities (inter-class variation) while treating disparate images of the same person as one identity (intra-class variation). However, the *degree of similarity* between faces of different identities is not explicitly modeled during the training process. Their work emphasizes the need for designing a system that handles face identity and face similarity differently.

B. Reranking

Computer vision tasks which involve the retrieval of a set of items, typically return results as an ordered list. The initial list is based on a set of general-purpose features that rank items based on their similarity and dissimilarity to the query (probe). The state-of-the-art for many problems in computer vision has not advanced to the stage such that the the rank-1 accuracy is perfect for all possible query images. Thus, a reranking technique, based on additional features to improve retrieval performance, is useful for computer vision tasks.

A number of reranking techniques have been proposed for a variety of problems (*e.g.*, object retrieval or person re-identification). Shen *et al.* propose a new similarity metric for object retrieval and enhance performance by reranking the initial search results using the nearest neighbors [19]. Garcia *et al.* propose a reranking method for person re-identification where small ambiguities specific to the top ranked matches are removed and the ambiguity-free images

are reranked [20]. Kim *et al.* compare the nearest neighbors of a probe and gallery to rerank a candidate list for person re-identification [21].

III. Match-Vicinity Plot

A goal of this work is to develop an adaptive selection scheme that selects a subset of samples in the ranked match list of an identification system for reranking. The criteria for the subset is two-fold: (1) the subset should be as small as possible and (2) the subset should include the gallery sample which matches the probe. These two criteria are inversely related: the smallest subset may exclude the correct gallery match, and a subset which guarantees the inclusion of the correct match has to necessarily include the entire gallery.

In this section, we analyze the distance scores in the ranked match list that may offer clues for selecting a subset of gallery samples that satisfies both criteria. Previous work has shown the benefit of analyzing the match scores in a ranked list. For example, Marasco *et al.* use the ratio of scores in the ranked match list to the rank-1 score as a feature vector and train a classifier to distinguish between correct and incorrect rank-1 matches [22].

Given a set of gallery images, $\mathcal{G} = \{g_1, g_2, \ldots, g_n\}$ and a probe image p, we compare the probe to each gallery image $g_i \in \mathcal{G}$, resulting in a set of distance scores $\{d_i\}$. The ranked match list is constructed such that the gallery is ordered by distance scores from the smallest to largest. This results in a ranked match list corresponding to the probe p:

$$\mathcal{L} = \left(d_p^{(1)}, d_p^{(2)}, \ldots, d_p^{(n)}\right),$$

where, $d_p^{(i)}$ is the i^{th} smallest distance score. Suppose the correct match occurs at rank c. A distance score at rank i can be normalized relative to the correct match score as, $s_p^{(i)} = d_p^{(i)} - d_p^{(c)}$.

The match scores in the vicinity of the correct match for a probe p can then be given by match-vicinity vector,

$$\begin{aligned}\phi_p &= \begin{bmatrix} s_p^{(c-5)} & \cdots & s_p^{(c-1)} & s_p^{(c)} & s_p^{(c+1)} & \cdots & s_p^{(c+5)} \end{bmatrix} \\ &= \begin{bmatrix} s_p^{(c-5)} & \cdots & s_p^{(c-1)} & 0 & s_p^{(c+1)} & \cdots & s_p^{(c+5)} \end{bmatrix},\end{aligned}$$

where, $s_p^{(c)} = d_p^{(c)} - d_p^{(c)} = 0$.

We use a *match-vicinity plot* to depict how the scores change before and after the correct match is encounted on the ranked match list. The x-axis of the match-vicinity plot reports the location relative to the correct match ($c \pm 5$) and the y-axis reports the normalized score ($s^{(i)}$). The plot depicts the mean and standard deviation of the normalized scores across m probes, that is, the mean and standard deviation for each column in the matrix,

$$\begin{bmatrix} \phi_{p_1} \\ \vdots \\ \phi_{p_m} \end{bmatrix} = \begin{bmatrix} s_{p_1}^{(c-5)} & \cdots & s_{p_1}^{(c-1)} & 0 & s_{p_1}^{(c+1)} & \cdots & s_{p_1}^{(c+5)} \\ & \vdots & & \vdots & & \vdots & \\ s_{p_m}^{(c-5)} & \cdots & s_{p_m}^{(c-1)} & 0 & s_{p_m}^{(c+1)} & \cdots & s_{p_m}^{(c+5)} \end{bmatrix}$$

where, p_i is the i^{th} probe sample.

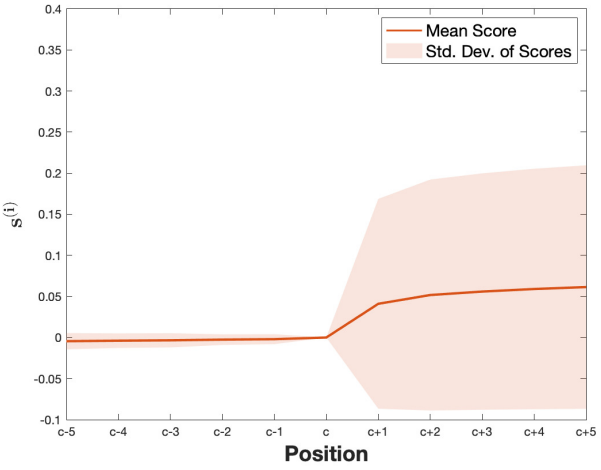


Fig. 3. Match-vicinity plot for the entire TinyFace dataset. The y-axis plots the mean and standard deviation of the normalized match score, $s^{(i)} = d^{(i)} - d^{(c)}$. The top correct match always occurs at position c (thus the y-axis value at rank c will always be 0 with a standard deviation of 0). The x-axis varies the rank $\pm$ 5 from rank c. Thus, the value at position $c-5$ is the mean and standard deviation of the normalized match score of the 5th item *before* the top correct match for all probe searches. Similarly, the value at rank $c+3$ is the mean and standard deviation of the normalized match score of the 3rd item *after* the top correct match for all probe searches.

Figure 3 shows the match-vicinity plot using the probe and gallery-match subsets of the TinyFace dataset (a further description of the dataset is given in Section VI). The probes are compared to each gallery sample using the ArcFace matcher (further information is given in Section V) which outputs a distance score. The ranked match list is tabulated by sorting the gallery samples in ascending order based on the distance score (*i.e.*, rank 1 is the gallery sample with the smallest distance score). As an example, consider two probe searches, the first with the top correct match at rank 10 and the second with the top correct match at rank 20. In the match-vicinity plot, we consider the samples at positions 5 to 15 for the first probe search and the samples at position 15 to 25 for the second probe search. The match-vicinity plot in Figure 3 indicates that the distance scores increase more rapidly *after* the correct match is found in the ranked gallery list.

We will exploit the observations above to develop a rerank selection scheme. We describe such a scheme in Section IV-B. The goal of an adaptive selection scheme is to select the top k items on the ranked match list where k is very close to, but not below, the rank of the correct match.

IV. Rerank Selection Scheme

In this work, we consider two schemes for selecting images in the ranked list that have to be reranked: (1) a fixed selection scheme and (2) an adaptive selection scheme.

A. Fixed Selection Scheme

The fixed selection scheme simply selects the top k ranks from the ranked match list. It is independent of the match

scores associated with each gallery sample. Thus, it is susceptible to unusually high correct match ranks. It will only consider the first k ranks and will simply ignore the correct match if it were to occur at a rank beyond k.

B. Adaptive Selection Scheme

The adaptive selection scheme determines the subset to be reranked based on the distance scores generated when comparing the probe with each gallery sample. Given a set of gallery images $\mathcal{G} = \{g_1, g_2, \ldots, g_n\}$ and probe image p, we compare the probe, p, to each gallery image $g_i \in \mathcal{G}$, resulting in a distance score d_i. The ranked match list is constructed such that the gallery is ordered by distance scores from smallest to largest. This results in a ranked match list,

$$\mathcal{L} = \left(d^{(1)}, d^{(2)}, \ldots, d^{(n)}\right),$$

where $d^{(i)}$ is the i^{th} smallest distance score.

In the ideal case, the adaptive selection scheme should select the top k matches where the correct match occurs at rank k. In Section III, the match-vicinity plot shows the distance scores increase at a higher rate from one rank to the next *after* encountering the correct match. This motivates us to consider a rolling sum to capture this phenomenon and select a rerank subset that is as small as possible but still potentially includes the correct match.

A rolling sum over q consecutive distance scores is tabulated over the ranked match list. This is given as,

$$S_k = \sum_{i=0}^{q-1} d^{(k-i)},$$

where $k \geq q$. The rerank subset is the first k matches in the ranked match list such that $S_k > \tau$ and k is minimized (*i.e.*, the smallest value of k that satisfies $S_k > \tau$).

V. Face Matching

This work makes use of two face matchers: (1) a general-purpose (GP) matcher and (2) a lookalike disambiguator (LD). The GP matcher is an existing, publicly-available matcher while the LD is adapted from the GP matcher for disambiguation of lookalike face images. Figure 4 shows an overview of how the LD is created. The lookalike faces can vary across different face matchers. But the proposed method can be used with any general-purpose face matcher.

A. General-Purpose (GP) Matcher

The GP matcher is an existing, publicly-available matcher. The ArcFace matcher [23] is selected due to it high performance (99.8% accuracy on the LFW datset [24]). The network, represented by $f(\cdot)$, takes an input image I and outputs a 512-dimensional representation of the face image (*i.e.*, $f(I) \in \mathbb{R}^{512}$). Two faces are compared using the euclidean distance metric, which yields a distance score between the two faces.

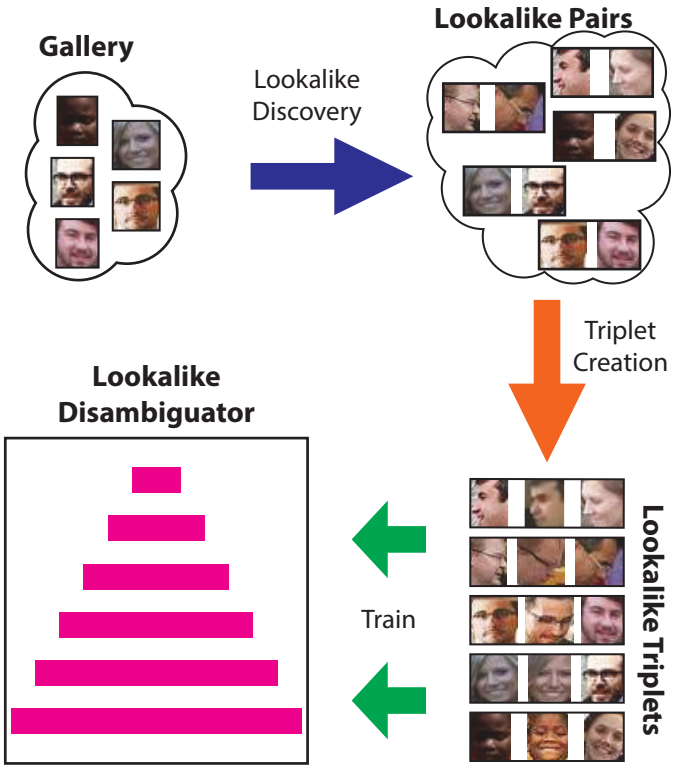


Fig. 4. Overview of the lookalike disambiguator creation. First, lookalike pairs are discovered from the gallery. Second, the lookalike pairs are converted to lookalike triplets with an anchor, positive, and negative sample. Lastly, the lookalike disambiguator network is trained using the lookalike triplets.

B. Lookalike Disambiguator (LD)

The Lookalike Disambiguator (LD) uses the same pre-trained network as the GP matcher. In our work, the network is further fine-tuned using lookalike triplets. A lookalike triplet consists of three images, (I_a, I_p, I_n), where I_a is the anchor sample, I_p is the positive sample, and I_n is the negative sample. Images I_a and I_p originate from the same subject, while images I_a and I_n originate from different subjects. The images I_a and I_n are a lookalike face pair (*i.e.*, they are an imposter pair such that the GP matcher predicts a small distance score between them).

The triplet embedding loss function is given by:

$$L = \sum_{\{(I_a, I_p, I_n)\}} \|f(I_a) - f(I_p)\|_2 - \|f(I_a) - f(I_n)\|_2 + \alpha_{\text{margin}}$$

where, $\|\cdot\|_2$ is the euclidean distance and α_{margin} is a user-tunable parameter for the margin of minimum distance between positive and negative samples. The network is fined-tuned in the PyTorch environment with $\alpha_{\text{margin}} = 0.2$ and a batch size of 32. Stochastic gradient descent trains the network with an Adam optimizer and learning rate of 0.01.

VI. Data and Lookalike Discovery

In this work, the TinyFace dataset is used as previous work shows this dataset to be a challenging one for the task of

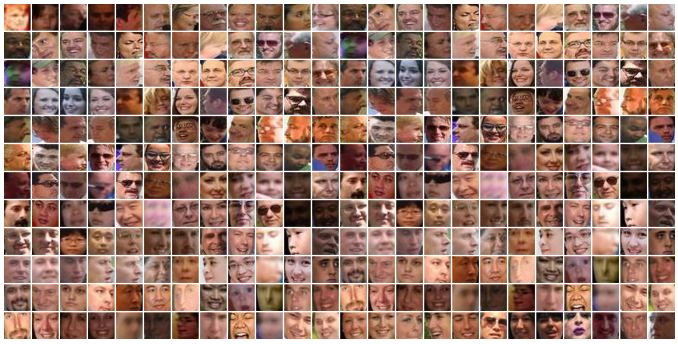


Fig. 5. Examples of face images in the TinyFace dataset [3].

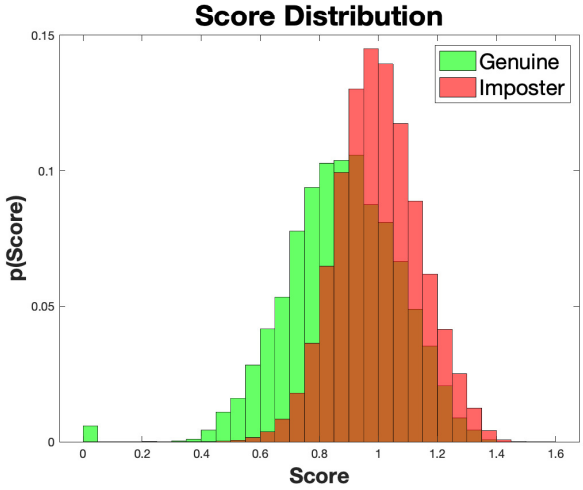


Fig. 6. Genuine and imposter score distribution of the entire TinyFace gallery set.

identification [3], [25]. The dataset is mined for lookalike images using the ArcFace matcher.

A. TinyFace Dataset

The TinyFace dataset [3] consists of small face images (average size 20×16 pixels). The test set comprises three subsets (probe, gallery-match, and gallery-distractor), of which, only two are used in this work (probe and gallery-match). 3,728 face images of 2,569 subjects compose the probe set while 4,443 face images of the same 2,569 subjects compose the gallery-match set (known as the gallery in this work since the gallery-distractor set is not used because there are no subject labels required for lookalike discovery). Figure 5 shows examples of the face images in the dataset. A filtered version of the probe and gallery-match subsets is also considered in this work. The filtered version of each subset is manually filtered to remove faces with a profile view. The filtered portion of the probe subset contains 2,081 images and the filtered portion of the gallery-match subset contains 2,461 images. The filtered version contains 1,145 subjects.

Fig. 7. Examples of lookalike pairs detected by the ArcFace matcher.

B. Lookalike Discovery

The lookalike disambiguator (LD) should be able to successfully distinguish between faces that lookalike. Lookalikes can be curated in many different ways. For example, a human annotator could review face image repositories, including the Web, in order to identify doppelgängers. Lookalikes could also be deduced using a face matcher. In this work, lookalike face images are chosen from the TinyFace dataset using the ArcFace matcher. This is preferable to human annotations as humans and machines may perceive faces differently, and this work focuses on *automated* face matchers.

The entire TinyFace *gallery* set is mined for lookalike face images by identifying imposter pairs with low distance scores. Figure 6 plots the genuine and imposter score distributions of the gallery set using the ArcFace matcher (the gallery contains multiple images of the same subject). We select imposter pairs whose distance score is less than 0.8 since $[0, 0.8]$ represents a range where $p(\text{score}|\text{genuine}) \gg p(\text{score}|\text{imposter})$. This results in the detection of ~679K lookalike pairs (6.9% of all imposter pairs). Figure 7 shows examples of detected lookalike face pairs.

Each lookalike pair is further augmented with additional images of the subjects in the pair to generate several lookalike triplets. A triplet consist of an anchor sample, a negative sample, and a positive sample. A lookalike pair comprises two images, *i.e.*, (I_i^1, I_j^1) where I_i^1 is the first image of subject i and I_j^1 is the first image of subject j. At least two triplets are constructed from a single lookalike pair. The first triplet, (I_i^1, I_i^2, I_j^1), contains the first image of subject i (I_i^1) as the anchor sample, the second image of subject i (I_i^2) as the positive sample, and the first image of subject j (I_j^1) as the negative sample. The second triplet, (I_j^1, I_j^2, I_i^1), contains the first image of subject j (I_j^1) as the anchor sample, the second image of subject j (I_j^2) as the positive sample, and the first image of subject i (I_i^1) as the negative sample. If subject i or j has more than two face samples each, then additional lookalike triplets can be constructed. The triplets are used to fine-tune the LD as described in Section V-B for 5 epochs.

VII. Experiments

Experiments include an evaluation of: (1) the proposed reranking selection scheme, and (2) an identification performance analysis with and without reranking the gallery samples using the Lookalike Disambiguator. Results are reported on the the filtered version of the TinyFace dataset as the identification results on the entire dataset is lower (due to pose variations). **As pose variation is not studied in this work, we use the filtered data which has profile-view faces removed.**

TABLE I
RESULTS OF PARAMETER SEARCH FOR q AND τ ON THE FILTERED TINYFACE GALLERY.

q	Surplus Size		Hit Rate	τ
	Total	Per Search		
1	270,276	142.5	55.77%	0.7695
2	294,003	155.0	61.68%	1.378
3	295,173	155.6	62.20%	1.958
4	296,353	156.2	62.63%	2.511
5	297,541	156.8	63.05%	3.049
6	298,737	157.5	63.52%	3.574
7	299,942	158.1	63.78%	4.090
8	301,152	158.8	63.94%	4.597
9	302,365	159.4	64.21%	5.094
10	303,583	160.0	64.63%	5.584

A. Parameter Selection

The adaptive selection scheme requires two parameters, τ (rolling sum threshold) and q (number of scores considered in rolling sum). These values are estimated from the gallery set. The gallery contains 2,461 face images of 1,145 subjects; some subjects have multiple images. Simulation of a probe search using the GP matcher (without using the probe set from the TinyFace data) is achieved by using one image from the gallery as a probe. This is repeated for each image in the gallery set for which at least one other image of the same subject exists (there are 1,897 such images). The rolling sum is tabulated for each search operation. The average value of the rolling sum across all searches is taken at the position where the top correct match occurs. This is repeated for varying values of q from 1 to 10.

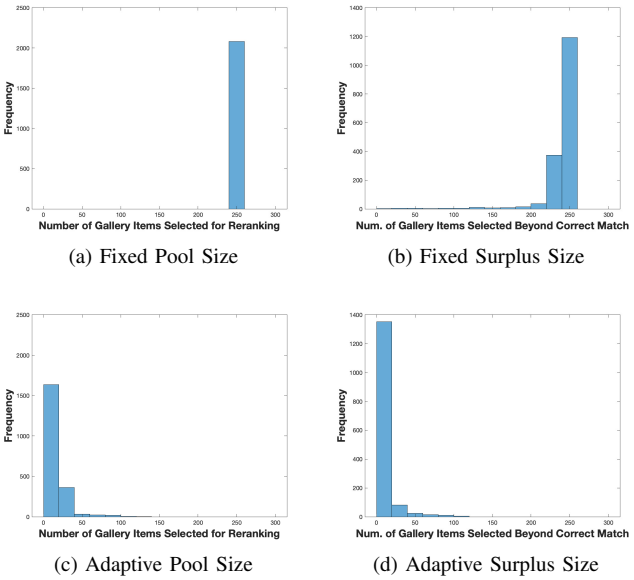
A perfect selection scheme would choose the minimum number of gallery items necessary to achieve 100% identification accuracy at rank 1. This means the scheme would choose the first k gallery items only if the correct gallery match for a probe occurs at rank k. This gives rise to two conflicting criteria: (1) select as few samples as possible and (2) ensure the selection includes the correct gallery match (the only way of ensuring this is to select the entire gallery). Thus, the efficacy of the selection for reranking is evaluated using two metrics, *hit rate* and *surplus size*. The hit rate measure the fraction of searches for which the selection scheme chooses a gallery subset that includes the correct match. The surplus size reports the number of samples included in the subset with rank higher than the rank of the correct match (*e.g.*, if the selected subset is of size 12 and the correct match occurs at rank 10, then this results in a surplus size of 2). The metrics are reported in Table I for q ranging from 1 to 10.

B. Rerank Selection Schemes

Two selection scheme are evaluated: fixed and adaptive. For each scheme, two histograms are plotted: (1) number of gallery items selected (*pool size*) and (2) number of gallery item selected beyond the correct match (*surplus size*). A small rerank pool size is generally better. However, a large pool size is not inherently bad: it could be that the correct match occurs at a lower rank so it is preferable to select a large number of gallery samples to rerank. This is why the surplus size is a useful metric. This plot show how many extra gallery samples at ranks beyond the correct match are selected as part of the subset to rerank.

TABLE II
POOL SIZE AND HIT RATE FOR FIXED AND ADAPTIVE RERANK SELECTION SCHEMES. THE POOL SIZE STATISTICS ARE MINIMUM/AVERAGE/MEDIAN/MAXIMUM.

Scheme	Pool Size	Hit Rate
Fixed	246/246/246/246	80.1%
Adaptive	15/20.66/18/121	71.3%



(a) Fixed Pool Size (b) Fixed Surplus Size

(c) Adaptive Pool Size (d) Adaptive Surplus Size

Fig. 8. Pool and surplus size for the fixed and adaptive rerank selection scheme.

In this experiment, $k = 246$ (top 10% of the ranked match list) for the fixed scheme and the parameters for the adaptive schemes are set to $\tau = 5.584$ and $q = 10$ (parameters with highest hit rate in Table I). Table II reports statistics about the number of gallery items selected and the hit rate. Figure 8 shows histograms of the pool size and the surplus size for both the fixed and adaptive schemes.

C. Identification with Reranking

This experiment evaluates the closed-set identification performance before and after reranking. The GP matcher compares the TinyFace gallery and probe sets to obtain a ranked match list. A subset of the ranked match list is reranked using the Lookalike Disambiguator. A Cumulative Match Characteristic (CMC) curve is tabulated from the ranked match list by recording what percentage of probes have the correct gallery match occur at rank i or better. Figure 9 shows the CMC curve before reranking (original ranking) and after reranking using the fixed and adaptive selection schemes.

VIII. ANALYSIS

Based on Table II, the fixed selection scheme appears more successful than the adaptive selection scheme as the hit rate

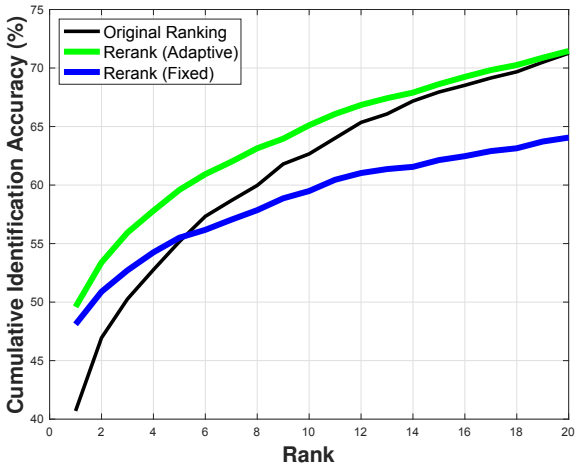
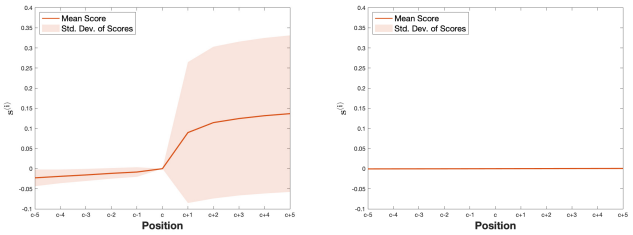


Fig. 9. Cumulative Match Characteristic (CMC) curve showing the identification accuracy at ranks 1 to 20. Note that both the proposed reranking schemes increase rank-1 accuracy. However, the adaptive scheme has a much smaller surplus size making it more efficient in terms of time and computation. Note that no alignment module was used by the face matcher.



(a) Ranks in Top 20 (b) Rank 500 and Beyond

Fig. 10. Match-vicinity plots for specific groups of probes in the filtered TinyFace dataset. (a) shows the match-vicinity plot only of probes for which the correct gallery samples occur in the top 20 ranks. red(b) shows the match-vicinity plot only of probes for which the correct gallery samples occur at rank 500 or higher.

is 80.1% (an increase of 8.8% over the adaptive scheme). However, the surplus size for the adaptive scheme is much smaller (compare Figure 8b and Figure 8d). In the fixed scheme, unlike the adaptive scheme, there is no mechanism to deduce if fewer than a predetermined number of samples are needed for a particular query. Thus, the adaptive scheme is able to select subsets for reranking with smaller surplus sizes than the fixed scheme.

Figure 10 shows two match-vicinity plots, one in which only searches where the correct match occurs in the top 20 ranks are included and another in which only searches where the correct match occurs at rank 500 or above are included. We see that in the case of probes with a "top 20" hit, there is a pronounced increase in the distance score after the correct match. Such an increase is conspicuously absent in the case of a "beyond 500" hit. This suggest that the adaptive scheme will work best for searches where the correct gallery sample occurs at higher ranks. This may also explain why the adaptive scheme has a lower hit rate than the fixed scheme (Table II). This means, the criteria used by the adaptive scheme to perform reranking occurs for a smaller proportion of probe searches.

The adaptive scheme also includes a smaller surplus size overall. The total surplus size across all probe searches for the fixed scheme is 393,538 versus 25,364 for the adaptive scheme – a reduction of 368,174 samples to be reranked (a 93.6% reduction). This is a reduction of 176.9 samples per probe query. The role of the lookalike disambiguator is to distinguish between similar-looking faces, a difficult problem. By reducing the number of samples that must be disambiguated, the chances of improving identification performance increases.

Figure 9 shows the identification accuracy for the original ranking by the GP matcher and the reranked performance using the fixed scheme and the adaptive scheme. The reranking improves performance modestly (an increase in identification accuracy of 7.40% at rank 1 for the fixed scheme and 8.89% at rank 1 for the adaptive scheme). The fixed and the adaptive schemes performed similarly for identification at the top few ranks (where improvement is most prominent), but the fixed scheme drops below the original ranking at rank 6 while the adpative scheme outperforms the original ranking. In addition, the adaptive scheme selects far fewer gallery samples for reranking making it more efficient. The hit rate for the fixed scheme is 8.8% higher than the adaptive scheme. This means that there are more rerank subsets selected by the fixed scheme which include the correct gallery match, but the rank-1 identification accuracy is still commensurate with that of the adaptive scheme. This suggests that the adaptive scheme only selects subsets with a chance for improved identification accuracy.

IX. Conclusion

This work addresses the lookalike problem for face identification systems. It proposes the use of a lookalike disambiguator to distinguish between similar-looking face images. This is achieved by selecting a subset of gallery images to rerank based on an initial ranking by a general-purpose face matcher. The selection scheme and lookalike disambiguator proposed in this work show a modest improvement in identification accuracy in a closed-set identification experiment.

Possible future work includes evaluating other features useful for lookalike disambiguation such as motion information. The subset selection scheme could possibly benefit from a principled approach for deducing the subset to rerank rather than rely on heuristic measures. This evaluation could be extended to other datasets and modalities to verify that the observations described in this work are applicable in other situations.

Acknowledgment

This work was partially funded by NSF CITeR.

REFERENCES

[1] A. K. Jain, A. A. Ross, and K. Nandakumar, *Introduction to Biometrics*. Springer, 2011.

[2] A. Sadovnik, W. Gharbi, T. Vu, and A. Gallagher, "Finding your lookalike: Measuring face similarity rather than face identity," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018, pp. 2345–2353.

[3] Z. Cheng, X. Zhu, and S. Gong, "Low-resolution face recognition," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2018, pp. 605–621.

[4] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. S. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer, "An introduction to the good, the bad, & the ugly face recognition challenge problem," in *Proceedings of the Automatic Face Gesture Recognition (FG)*, 2011.

[5] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2013, pp. 113–120.

[6] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *Proceedings of the International Conference on Computer Vision (CVPR)*, 2015, pp. 676–684.

[7] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the International Conference on Computer Vision (CVPR)*, Honolulu, HI, July 2017.

[8] Z. Sun, A. A. Paulino, J. Feng, Z. Chai, T. Tan, and A. K. Jain, "A study of multibiometric traits of identical twins," in *Proceedings of SPIE Defense, Security, and Sensing: Biometric Technology for Human Identification*, 2010.

[9] K. W. Bowyer, "What surprises do identical twins have for identity science?" *Computer*, vol. 44, no. 7, pp. 100–102, 2011.

[10] B. Klare, A. A. Paulino, and A. K. Jain, "Analysis of facial features in identical twins," in *Proceedings of the International Joint Conference on Biometrics (IJCB)*, Washington, DC, 2011, pp. 1–8.

[11] P. J. Phillips, P. J. Flynn, K. W. Bowyer, R. W. V. Bruegge, P. J. Grother, G. W. Quinn, and M. Pruitt, "Distinguishing identical twins by face recognition," in *Proceedings of the Automatic Face Gesture Recognition (FG)*, 2011, pp. 185–192.

[12] K. W. Bowyer and P. J. Flynn, "Biometric identification of identical twins: A survey," in *Proceedings of International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Niagara Falls, NY, 2016, pp. 1–8.

[20] J. Garcia, N. Martinel, A. Gardel, I. Bravo, G. L. Foresti, and C. Micheloni, "Discriminant context information analysis for post-ranking person re-identification," *Transactions on Image Processing*, vol. 26, no. 4, pp. 1650–1665, 2017.

[13] N. Srinivas, G. Aggarwal, P. J. Flynn, and R. W. Vorder Bruegge, "Analysis of facial marks to distinguish between identical twins," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 7, no. 5, pp. 1536–1550, 2012.

[14] T. H. N. Le, K. Luu, K. Seshadri, and M. Savvides, "A facial aging approach to identification of identical twins," in *Proceedings of International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Arlington, VA, 2012, pp. 91–98.

[15] X. Sun, A. Torfi, and N. Nasrabadi, "Deep siamese convolutional neural networks for identical twins and look-alike identification," in *Deep Learning in Biometrics*, M. Vatsa, R. Singh, and A. Majumdar, Eds. Boca Raton: CRC Press, 2018, ch. 3, pp. 65–83.

[16] H. Lamba, A. Sarkar, M. Vatsa, R. Singh, and A. Noore, "Face recognition for look-alikes: A preliminary study," in *Proceedings of the International Joint Conference on Biometrics (IJCB)*, Washington, DC, 2011, pp. 1–6.

[17] E. Smirnov, A. Melnikov, S. Novoselov, E. Luckyanets, and G. Lavrentyeva, "Doppelganger mining for face representation learning," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (CVPRW)*, 2017, pp. 1916–1923.

[18] A. Moeini, K. Faez, H. Moeini, and A. M. Safai, "Open-set face recognition across look-alike faces in real-world scenarios," *Image and Vision Computing*, vol. 57, pp. 1–14, 2017.

[19] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu, "Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking," in *Proceedings of the International Conference on Computer Vision (CVPR)*, 2012, pp. 3013–3020.

[21] K. Kim, M. Byeon, and J. Y. Choi, "Re-ranking with ranking-reflected similarity for person re-identification," *Pattern Recognition Letters*, vol. 128, pp. 326–332, 2019.

[22] E. Marasco, A. Ross, and C. Sansone, "Predicting identification errors in a multibiometric system based on ranks and scores," in *Proceedings of the International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2010, pp. 1–6.

[23] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[24] 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," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[25] Y. Shi, X. Yu, K. Sohn, M. Chandraker, and A. K. Jain, "Towards universal representation learning for deep face recognition," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6817–6826.