Toward Open-Set Face Recognition

Manuel Günther   Steve Cruz   Ethan M. Rudd

Terrance E. Boult

Vision and Security Technology Lab, University of Colorado Colorado Springs

{mgunther,scruz,erudd,tboult}@vast.uccs.edu

Abstract

Much research has been conducted on both face iden-tification and face verification, with greater focus on the latter. Research on face identification has mostly focused on using closed-set protocols, which assume that all probe images used in evaluation contain identities of subjects that are enrolled in the gallery. Real systems, however, where only a fraction of probe sample identities are enrolled in the gallery, cannot make this closed-set assumption. Instead, they must assume an open set of probe samples and be able to reject/ignore those that correspond to unknown identities. In this paper, we address the widespread misconception that thresholding verification-like scores is a good way to solve the open-set face identification problem, by formulating an open-set face identification protocol and evaluating differ-ent strategies for assessing similarity. Our open-set identi-fication protocol is based on the canonical labeled faces in the wild (LFW) dataset. Additionally to the known identi-ties, we introduce the concepts of known unknowns (known, but uninteresting persons) and unknown unknowns (people never seen before) to the biometric community. We compare three algorithms for assessing similarity in a deep feature space under an open-set protocol: thresholded verification-like scores, linear discriminant analysis (LDA) scores, and an extreme value machine (EVM) probabilities. Our find-ings suggest that thresholding EVM probabilities, which are open-set by design, outperforms thresholding verification-like scores.

1. Introduction

Face recognition algorithms have been widely re-searched over the past decades, resulting in tremendous performance improvements, particularly over the past few years. Even traditional face recognition algorithms, i.e., before the widespread use of deep networks, performed quite well on frontal images under good illumination [12], making them commercially viable for certain applications. For instance, verification scenarios such as automated bor-
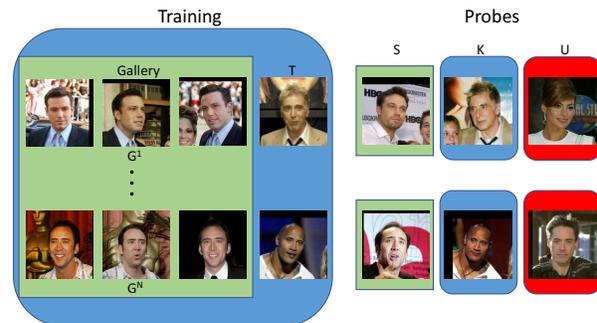

Figure 1: OPEN-SET RECOGNITION. *Closed-set identi-fication performs comparisons between the Gallery and known probe (S)ubjects. The open-set identification protocol presented in this paper requires additional more subtle comparisons, due to the presence of known unknowns (uninteresting subjects) in the (T)raining set, (K)nown unknown probes of the same identity at query time, and (U)nknown unknowns whose identities are only seen during query time. The open-set identification objective is to correctly identify probe (S)ubjects that are present in the gallery while rejecting all other probe queries as unknown.*

der control stations [7] allow reasonable control of imaging conditions and subjects usually cooperate with the system – those that do not are easily spotted by airport security per-sonnel. As of 2006, O'Toole *et al.* [23] demonstrated that algorithmic solutions were able to outperform humans for such constrained recognition tasks. Therefore, researchers have shifted their focus to more difficult conditions.

Günther *et al.* [12] have found that these traditional face recognition algorithms are not designed to and, therefore, do not perform very well on images with uncontrolled fac-tors such as facial expression, non-frontal illumination, par-tial occlusions of the face, or non-frontal face pose, which occur in modern face recognition datasets [15, 18]. While different strategies have been proposed to improve the per-formance of traditional algorithms across pose, e.g., using face frontalization techniques [14] or 3D modeling [16], the introduction of deep convolutional neural networks (DC-NNs) for face recognition [36, 24] has overcome the pose issue to a significant extent. For example, deep neural net-

works have outperformed traditional methods by such a wide margin on the labeled faces in the wild (LFW) benchmark [15] that this once challenging benchmark is now considered quite easy, at least under the conventional verification protocol. With these improved representations, face recognition based on DCNNs can now, theoretically, be used in more complicated scenarios, e.g., to identify criminals in surveillance camera images.

However, the identification problem introduces new and different challenges compared to the verification scenario. While verification requires only a single $1:1$ comparison, identification requires $1:N$ comparisons – of a probe sample with templates from many subjects in a gallery, e.g., the watch-list of criminals. Depending on the gallery size N, finding the correct identity can be a much harder task than simply performing a correct $1:1$ comparison. Furthermore, in many scenarios the gallery can combine multiple images of each enrolled subject to build a more effective template. An even more subtle aspect of the identification problem is that, in many scenarios, most subjects in probe images are not contained in the gallery at all. Going back to our previous example, we would hope that the majority of the population is not present in a criminal database. Thus, the recognition algorithm should be able to work under an $M:N$ open-set protocol, with the ability to detect and ignore probe samples with identities that are not present in the gallery. This is achieved by giving low similarity values to all known subjects in the gallery, which is a more intricate task than one would expect. Contrary to some people's intuition, good open-set recognition is not as simple as thresholding a $1:1$ verification approach on raw similarity scores. Although this thresholding does reject some unknowns, we show that there are more effective techniques.

Open-set recognition is clearly desirable for many biometric recognition systems, particularly face. For example, surveillance cameras in airports capture people and compare their faces with a watch-list of *known* criminals. The airport staff, which is not included in the watch-list but regularly passes through the eye of the camera, should not confuse the algorithm. Hence, this list of known, but uninteresting people can be seen as *known unknowns* during training. Finally, many *unknown unknowns*, i.e., passengers that are not on the watch-list and sojourn in the airport need to be ignored by the face recognition algorithm.

In this paper we introduce a small open-set face identification evaluation protocol based on the widely used LFW dataset, which previously has mainly been used for evaluating face verification systems. Particularly, we introduce *known unknowns*, i.e., probe images at query time with identities that were used during training but are not present in the gallery; and *unknown unknowns*, i.e., subjects at query time whose identities have never been seen by the system, neither during training nor during enrollment.

An illustration of these concepts is shown in Fig. 1. While the LFW dataset is now considered easy under a verification protocol, we show that under our open-set identification protocol for this dataset is still quite challenging. Note that this is the first protocol that deals with known and unknown unknowns; similar protocols can be generated for other datasets.

We evaluate three different approaches to open-set face recognition. Each of these algorithms operate on the same high-quality deep features, which we extract using the publicly available VGG face network [24]. First, we evaluate a standard $1:1$ verification-like technique that is applied to deep features [34, 6], i.e., we compute cosine similarities between deep features in gallery templates and probes; to reject unknowns we use a threshold on the similarity score. Second, we perform the standard linear discriminant analysis (LDA) technique [39] on top of the deep features, making use of the known unknowns during training. We then use the learnt projection matrix to project the original deep feature vectors to the LDA subspace and compare gallery templates to probes via cosine similarity. Finally, to model probabilities of inclusion with respect to the support of gallery samples, we train an extreme value machine (EVM) [26] on cosine distances between deep features, again using known unknowns during training. We show that the raw cosine similarity performs well in a closed set scenario but not in an open-set setting, LDA can detect known unknowns very well but not the unknown unknowns, while EVM can handle both open-set cases with similar precision.

2. Related Work

The need for open-set face recognition has been widely acknowledged for well over a decade [11, 25]. Nonetheless, only a few works, e.g., [35, 9, 33, 20, 4, 21] have addressed the problem by predominantly focusing on obtaining an ad hoc rejection threshold on similarity score under an open-set evaluation protocol [25]. For example, Best-Rowden *et al.* [4] showed that a simple thresholding of a commercial of the shelf (COTS) algorithm works perfectly for verification, but does not provide decent open-set identification performance. The development of classifiers that explicitly model probability of inclusion [26] of probe samples with respect to a region of known support of the gallery has received far less attention in the face recognition community. For security-oriented applications where the enrollment process must be quick, the cost of false alarms is high, and the cost of missed alarms is even higher, the notion of using an ad hoc rejection threshold on similarity is problematic because the concept of *unknown* may change as more samples are enrolled and data bandwidth is variable, so a *one size fits all* threshold may not work well. A classifier that can efficiently be retrained with each enrolled gallery template to autonomously assess the probability that probe

data comes from regions of known support on behalf of the gallery while considering variable data bandwidths is a far more appealing alternative.

Thus, the motivation for applying classifiers that are open-set-by-design to face recognition problems is manifested. Several such classifiers have been developed in the computer vision community [26, 31, 32, 3], but their application has been limited to toy problems on modifications of canonical computer vision datasets like MNIST [19], or to generic object recognition problems like the ImageNet challenge [27]. However, object recognition problems inherently differ from biometric applications insofar as they are far more coarse-grained, the notion of enrollment does not exist, and deep learning solutions can be obtained by training an end-to-end network on the training set and using that end-to-end network as a classifier.

Face identification systems that use deep features [24, 34, 36, 6], by contrast, use truncated forward passes over pre-trained networks to extract features at enrollment or query time. The networks are trained in an end-to-end manner on labeled face identities, which generally differ from the identities enrolled into gallery templates. Templates are constructed during enrollment, e.g., by collecting extracted feature vectors from several images of each given subject. At query time, probe templates consisting of one or more extracted feature vectors of one subject, are matched against gallery templates. The identification procedure commonly takes the form of finding the gallery template with the sample of maximum similarity to the corresponding probe. Cosine is a common measure of similarity between feature vectors extracted from a face network [34, 6]. Particularly, when templates vary in number of images, feature vectors are sometimes aggregated for a given identity prior to matching, e.g., by taking the mean feature vector [24, 6].

To further enhance the performance of a similarity measure between two feature vectors, a common technique is to learn a transformation matrix across the training set that increases the similarity of vectors from the same class, while decreasing the similarity of vectors from different classes. Linear discriminant analysis (LDA) [5], which learns a subspace projection that minimizes the ratio of intra-class to inter-class variance over the training set, is one of the fundamental techniques that has been applied to face recognition [39] in the past. More advanced techniques include joint Bayesian [34] and triplet-loss [24, 28] embeddings, but they require more training data.

While ad hoc thresholding of raw similarity measures between features and projections thereof can lead to improved open-set face recognition performance, in this paper we compare the results of using such techniques to using a lightweight classifier – the extreme value machine (EVM) introduced by Rudd *et al.* [26] – that is open-set by design.

EVM uses statistical extreme value theory (EVT) based calibrations over margin distributions to obtain a *probability of sample inclusion* of each probe sample with respect to a gallery template. In doing so, it implicitly accounts for varying data bandwidths, yielding superior bounds on open space to those of a raw thresholded similarity function.

EVM has some similarities to cohort normalization techniques [37], parts of it can be viewed as an improved way of zero normalization (Z-norm) [2]. However, while Z-norm assumes Gaussian distribution of the data and takes into account all cohort data points – even if they are far away from the gallery template – EVM only considers the points with the highest similarities, and fits a Weibull distribution on half the distance in order to model *margin* distributions. While Scheirer *et al.* [30] showed that EVT can be successfully applied for score normalization in score fusion of biometric algorithms, in this paper we investigate its application to build gallery templates for open-set recognition.

3. Approach

As many readers might not be familiar with open-set evaluation, let us first introduce our exemplary implementation of an open-set protocol and explain the evaluation in more detail, before we discuss the tested algorithms.

3.1. Open-Set Face Recognition Protocol

Open-set face recognition has not been studied, in part due to the dearth of open-set evaluation protocols for face databases. Although the IJB-A dataset [18] provides an open-set protocol, IJB-A has a lot of different issues such as missing annotations, many profile and low quality images, and huge template sizes for both enrollment and querying. Hence, all these issues have to be solved before researchers can tackle the open-set problem using this dataset.

Also, neither IJB-A nor any other publicly available face recognition dataset provides an evaluation protocol to test open-set identification with both known unknowns and unknown unknowns. For example, [9, 33] only tests known unknowns, while other protocols [21, 20, 4] have disjoint training and enrollment set, which only allows to test unknown unknowns. Thus, we implemented our own evaluation protocol, which is non-random, simple and can easily be implemented. We chose to generate an open-set protocol for the labeled faces in the wild (LFW) dataset [15], for several reasons. First, LFW is publicly available, well-investigated, and contains relatively unconstrained imaging conditions. Second, LFW is large enough to provide meaningful results, yet it is small enough that experiments can be run using a normal desktop computer. Finally, LFW contains several identities, for which only a single image is present – which fits perfectly into our open-set concept.

We have split the identities in the LFW dataset into three groups. Those 602 identities with more than three images

are considered to be the *known* population. The 1070 identities with two or three images are the *known unknowns*, while the 4096 identities with only one image are considered to be *unknown unknowns*. The training set $\mathbf{T}$ contains the first three images (i.e., the images ending with `0001.jpg`, `0002.jpg` and `0003.jpg`) for each of the known identities, and one image (the image ending with `0001.jpg`) for the known unknowns. The enrollment set $\mathbf{G}$ is composed of the same three images for each of the known identities, which makes the protocol *biased*. Note that there are no unknowns (neither known nor unknown unknowns) inside the enrollment set.

Finally, we created four different probe sets, C, O_1, O_2, and O_3. The closed-set C contains the remaining images $\mathbf{S}$ of the known subjects, where the number of probe images per identity can vary between 1 and 527 (i.e., for George W. Bush). This set is used to evaluate closed-set identification and verification. Probe set $O_1 = \mathbf{S} \cup \mathbf{K}$ contains the same images as in the closed probe set C, and additionally the images $\mathbf{K}$ from the known unknowns, which were not part of the training set, one or two images per identity. $O_2 = \mathbf{S} \cup \mathbf{U}$ contains the closed-set images of C and the unknown unknowns $\mathbf{U}$, one per identity, which have not been seen during training and enrollment. Finally, the probe set O_3 contains all probe images, including known subjects, known unknowns, and unknown unknowns, i.e., $O_3 = \mathbf{S} \cup \mathbf{K} \cup \mathbf{U}$.

3.2. Evaluation

The closed-set evaluation uses standard cumulative match characteristics (CMC) curves and receiver operating characteristic (ROC) curves. Open-set recognition uses the detection and identification rate (DIR) curves as proposed in the Handbook of Face Recognition [25].

Cumulative match characteristics curves plot the identification rate, a.k.a. the recognition rate, with respect to a given rank. For each known probe $P \in \mathbf{S}$ of identity p, the rank r is computed as the number of subjects g in the gallery that are more similar than the correct subject, i.e.:

$$\text{rank}(P) = \left| \left\{ G^g \mid s(G^g, P) \geq s(G^p, P); G^g \in \mathbf{G} \right\} \right| \quad (1)$$

for a given similarity function $s(\cdot, \cdot)$. This means that rank $r = 1$ is assigned when the correct subject is the most similar one. The CMC curve plots illustrate the relative number of probes that have reached *at least* rank r.

Detection and identification rate curves plot the identification rates with respect to the false alarm rates, which should not be confused with false acceptance rates in ROC curves. For a given similarity threshold θ, a false alarm is issued when the similarity of an unknown probe $P \in \mathbf{K} \cup \mathbf{U}$ to *any* of the gallery subjects is higher than θ. The false alarm rate computes the average probability of these [25]:

$$\text{FAR}(\theta) = \frac{\left| \left\{ P \mid \max_g s(G^g, P) \geq \theta;\ P \in \mathbf{K} \cup \mathbf{U} \right\} \right|}{|\mathbf{K} \cup \mathbf{U}|}, \quad (2)$$

while the detection and identification rate for a given rank r is calculated on the known probe set, and given by [25]:

$$\text{DIR}(\theta) = \frac{\left| \left\{ P \mid \text{rank}(P) \geq r \wedge s(G^p, P) \geq \theta; P \in \mathbf{S} \right\} \right|}{|\mathbf{S}|}. \quad (3)$$

When plotting the DIR curve, different values for the threshold θ can be computed based on a given false alarm rate ϑ. After sorting the scores from (2) descendantly:

$$\text{scores} = \text{sort}(\{\max_g s(G^g, P) \geq \theta \mid P \in \mathbf{K} \cup \mathbf{U}\}) \quad (4)$$

the threshold can be computed by taking the smallest score $\theta > \theta'$, where:

$$\theta' = \text{scores}\left[\lfloor \vartheta \cdot |\mathbf{K} \cup \mathbf{U}| \rfloor \right] \quad (5)$$

Note that the threshold θ does not exist when θ' is already the maximum score.

3.3. Compared Methods

3.3.1 Cosine Similarity

Most face recognition algorithms that work on deep features simply apply a cosine similarity between pairs of deep feature vectors. Thus, we obtain a baseline measurement by computing the cosine similarity between the deep feature vectors of gallery template G^g of subject g and probe P. Since each gallery template is composed of three deep feature vectors: $G^g = (G_0^g, G_1^g, G_2^g)$, we apply two strategies: First, we compute three similarities and take the maximum value, which has been shown to provide the best performance in handling several scores [12]:

$$s_{\max}(G^g, P) = \max_{i \in \{0,1,2\}} \cos\left(G_i^g, P\right), \quad (6)$$

and second, we average the three deep features vectors [6]:

$$\bar{G}^g = \frac{1}{3} \sum_{i \in \{0,1,2\}} G_i^g \quad (7)$$

and compute the similarity between this average and the probe feature vector:

$$s_{\text{avg}}(\bar{G}^g, P) = \cos\left(\bar{G}^g, P\right). \quad (8)$$

Without further processing, this similarity is used inside the evaluation.

3.3.2 Linear Discriminant Analysis

To introduce a learning algorithm that can make use of the known unknowns during training, we select linear discriminant analysis (LDA) to learn a projection matrix W. First, we compute a principal component analysis (PCA) projection matrix. After projecting all training features $T \in \mathbf{T}$

into the PCA subspace, we train the LDA with the 603 classes of the training set $\mathbf{T}$, i.e., one class for each of the 602 known gallery subject, and one class containing the known unknowns. For more details on how to train a PCA+LDA projection matrix, please refer to [39, 38]. Finally, we project all enrollment and probe features into the combined PCA+LDA subspace using projection matrix W:

$$y_{G_i^g} = W^T G_i^g, \quad y_{\bar{G}^g} = W^T \bar{G}^g, \quad y_P = W^T P. \quad (9)$$

Scores are computed using the functions introduced in (6) and (8) on the projected features:

$$s_{\max}(y_{G_i^g}, y_P) \qquad s_{\text{avg}}(y_{\bar{G}^g}, y_P). \quad (10)$$

3.4. Extreme Value Machine

For a third approach, we choose the extreme value machine (EVM) introduced by Rudd *et al.* [26]. While EVM was formulated to handle generic classification tasks, we utilize the algorithm to perform biometric identification. The EVM classifier uses statistical extreme value theory (EVT) [10] to perform nonlinear, kernel-free classification, optionally in an incremental learning setting. The classifier fits an EVT distribution per point over several of the nearest fractional radial distances to points from other classes, and uses a statistical rejection model on the resultant cumulative distribution function (CDF) to model probability of sample inclusion (PSI or Ψ). Taking a fixed number of the data point and distribution pairs per class that optimally summarize each class of interest yields a compact probabilistic representation of each class in terms of extreme vectors (EVs).

We tailor EVM to a face identification similarity function by letting each feature vector be associated with an identity. Deviating slightly from the original formulation, in which the fractional distance over which to fit EVT distributions was assumed to be $\alpha = 0.5$ times the distance (cf. (12)) to formalize the classifier in terms of fitting margin distributions, we formalize the distance multiplier in terms of hyperparameter α. With $\alpha \neq 0.5$, EVM no longer models maximum margin distributions, but rather a biased margin distribution. However, the margin distribution theorem from Rudd *et al.* [26], which governs the functional form of the EVT distribution for modeling probability of sample inclusion Ψ, still holds – dictating that the low tail of multiplied distances will follow a Weibull distribution. Applying a statistical rejection model to the resultant CDF, each feature vector within the gallery will have its own Ψ distribution. Denote the ith feature vector for gallery subject g as G_i^g. The resultant probability that probe P is associated with G_i^g is given by:

$$\Psi(G_i^g, P; \kappa_i^g, \lambda_i^g) = \exp^{-\left(\frac{d(G_i^g, P)}{\lambda_i^g}\right)^{\kappa_i^g}}, \quad (11)$$

where $d(G_i^g, P) = 1 - \cos(G_i^g, P)$ is the cosine distance of a probe P from a subject's feature G_i^g, and κ_i^g, λ_i^g are

Weibull shape and scale parameters, respectively. These parameters are obtained for each gallery feature G_i^g by computing all distances:

$$\text{dist} = \left\{\alpha \cdot d(G_i^g, T) \mid t \neq g; T \in \mathbf{T}\right\} \quad (12)$$

for all training set features $T \in \mathbf{T}$ with identity t, which do not correspond to the gallery identity g. A Weibull distribution is fit to the low tail of dist:

$$\text{dist}_\tau = \left\{d \mid d \in \text{dist} \wedge d < \theta^\tau\right\} \qquad \text{with} \quad (13)$$

$$\theta^\tau = \max_\theta \left|\left\{d \mid d \in \text{dist} \wedge d \leq \theta\right\}\right| = \tau, \quad (14)$$

where the tail size τ represents a second hyperparameter of EVM. For details on how to fit Weibull distributions on dist_τ, please refer to [26] or the MetaRecognition library.[1]

Another modification that we make to the original EVM algorithm is that we retain all EVs rather than select only the most informative ones. The main purpose of model reduction in [26] is to maintain compact representations in incremental learning settings. To be comparable to the other two algorithms, instead we perform scoring in two different ways, one where we compute the maximum probability over each feature G_i^g inside the gallery template G^g:

$$s_{\max}(G^g, P) = \max_{i \in \{0,1,2\}} \Psi(G_i^g, P, \kappa_i^g, \lambda_i^g). \quad (15)$$

For the other technique, we use the average $\bar{G}^g$ from (7) for each gallery template, compute distances between $\bar{G}^g$ and training set features T as in (12), and compute a Weibull fit on the tail of them in order to obtain $\bar{\kappa}^g$ and $\bar{\lambda}^g$. The final probability of inclusion is given as:

$$s_{\text{avg}}(G^g, P) = \Psi(\bar{G}^g, P, \bar{\kappa}^g, \bar{\lambda}^g). \quad (16)$$

4. Experiments

We conduct experiments on our novel open-set LFW protocol. For feature extraction, we use the VGG face network.[2] We employ the *Bob* signal processing library [1] to align the eye locations[3] of input images to fixed locations $(82, 112)$ and $(142, 112)$ pixels, to approximately mimic the alignment required for the VGG face network [24]. We use *Caffe* [17] to extract the 4096-dimensional `fc7` layer features from the VGG network, after removing all following layers from the network `prototxt`, including the `ReLU` layer. We perform EVT calibration using *libMR* [29]. Finally, to compute and plot the closed- and open-set evaluation results, *bob.measure*[4] is employed.

---

[1]The C and Python implementation of *libMR* is provided in:
http://pypi.python.org/pypi/libmr
[2]The *Caffe* model of the VGG network was downloaded from:
http://www.robots.ox.ac.uk/~vgg/software/vgg_face
[3]Annotations are provided under:
http://lear.inrialpes.fr/people/guillaumin/data.php
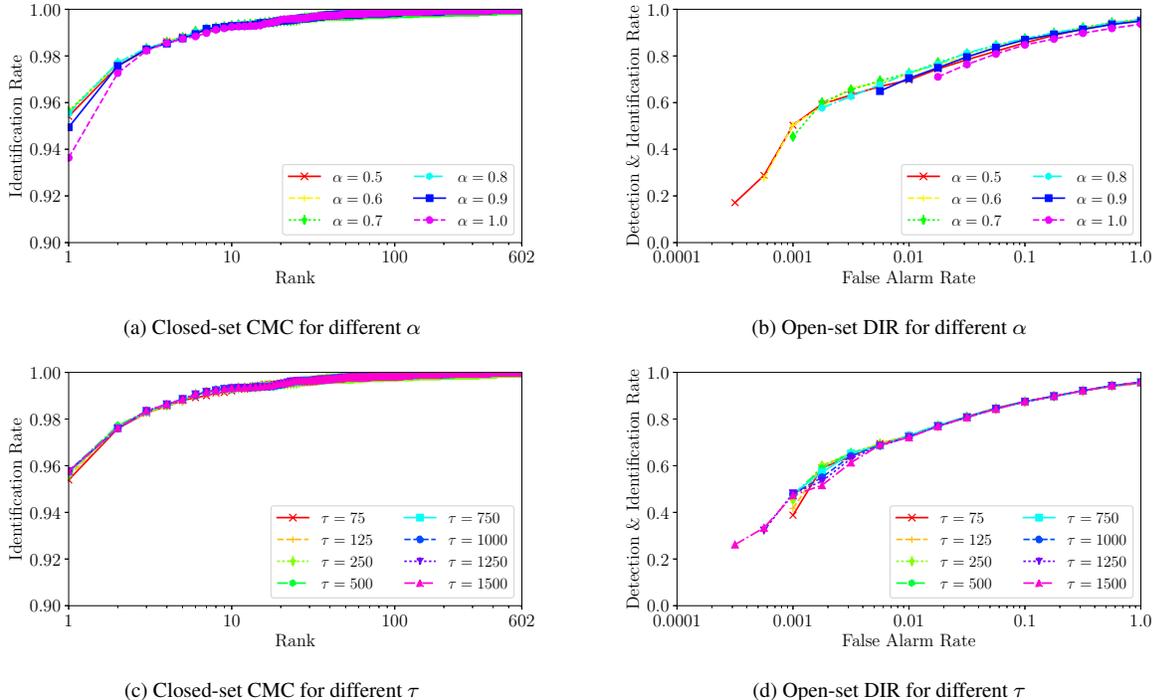[4]*Bob* and the *bob.measure* package can be found under:
http://www.idiap.ch/software/bob

Figure 2: EVM HYPERPARAMETER SELECTION. *Two different parameters of EVM are optimized: the distance multiplier α and the tail size τ. Both are evaluated using the CMC curve on probe set $\mathbb{C}$ and the DIR curve on probe set $\mathbb{O}_3$.*

4.1. Hyperparameter Selection

In the first set of experiments, we evaluate the effects of different hyperparameters for LDA and EVM. We only test different hyperparameters using the averaging approach, as experiments run faster with it, and we evaluate closed-set identification using probe set $\mathbb{C}$, and open-set recognition using the combined probe set $\mathbb{O}_3$. For LDA, only a single hyperparameter is optimized, which is the number of PCA components retained. After experimenting with several values, we find that retaining 99 % of the PCA variance gave the best results, leading to a final PCA+LDA subspace size of 256 dimensions.

EVM has two hyperparameters: the distance multiplier α and the tail size τ. We start optimizing α by setting the tail size to a hand-picked value of $\tau = 250$. Closed-set and open-set evaluation of different values of α are shown in Fig. 2(a) and 2(b), respectively. For closed-set, differences can only be seen in the very low ranks, after rank 5 all CMC curves seem to overlap completely. The best α value is between 0.6 and 0.8 and performance degrades slightly for smaller and larger values of α. Nevertheless, rank 1 identification rates are very high and do not vary substantially with any choice of α in the tested range. When examining the open-set performance in Fig. 2(b), we can see slightly larger differences for different values of α. For large α values, many unknown probes have the probability of 1 to belong to one of the gallery subjects and, hence, thresholds for

low false alarm rates cannot be computed. When the weight multiplier α gets lower, fewer of these cases occur. Based on both plots, we decide to use $\alpha = 0.7$ as a good trade-off between closed-set and open-set performance.

To evaluate different values for the tail size τ, we keep $\alpha = 0.7$ fixed. Examining the closed-set CMC curve in Fig. 2(c), we can again see very little difference. Generally, larger tail sizes seem to lead to better rank 1 identification rates, but already for rank 3, there is no apparent difference between any of the tested values. The open-set DIR curve given in Fig. 2(d) reveals that the open-set performance deteriorates for high and low tail sizes, while $\tau = 500$ seems to provide the best overall performance.

4.2. Comparison between Methods

After obtaining the optimal hyperparameters for LDA and EVM, we compare the performances of all three methods, and also with both scoring approaches, i.e., $s_{\max}$ and s_{avg}. The closed-set performance of Cos, LDA and EVM on probe set $\mathbb{C}$ is given in Fig. 3(a) and 3(b). As expected and reported [4], both closed-set identification and verification reach very high accuracies, e.g., a rank 1 identification rate of up to 96 %. One apparent observation is that the averaging strategy works better for all of the algorithms, which conforms with prior work [6] using deep features for face recognition. Interestingly, all algorithms perform almost similar in the closed-set identification task shown in
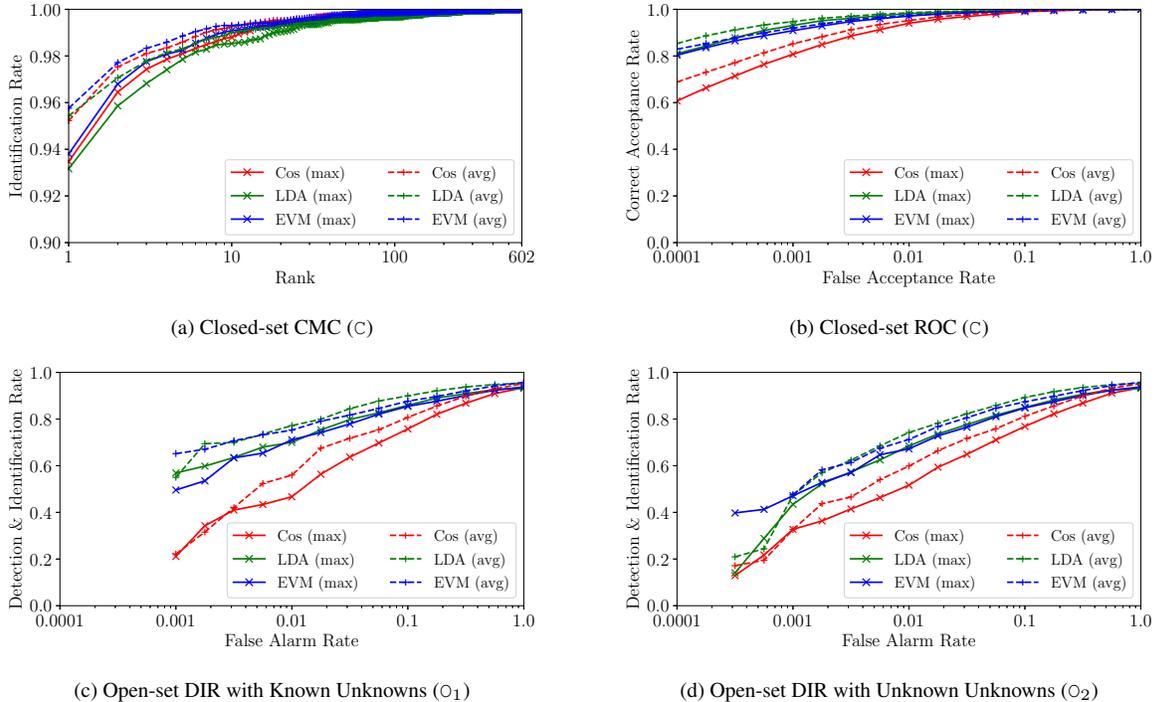
(a) Closed-set CMC (C)

(b) Closed-set ROC (C)

(c) Open-set DIR with Known Unknowns (O_1)

(d) Open-set DIR with Unknown Unknowns (O_2)

Figure 3: COMPARISON BETWEEN METHODS. *The (a) CMC curves and the (b) ROC curves for the closed-set evaluation on probe set C, as well as the open-set DIR curves for (c) probe set O_1 and (d) probe set O_2 are given for all six evaluated methods.*

Fig. 3(a), while EVM has a slight advantage over Cos, and LDA performs worst. However, when evaluating verification performance in Fig. 3(b), the simple cosine distance seems not to be as good as either LDA or EVM, indicating that distances are distributed differently for different identities (consisting with [8]), but both EVM and LDA effectively normalize them out.

More interestingly, looking at the open-set performance in Fig. 3(c), where we evaluate the known unknowns of probe set O_1, it is obvious that Cos has lost much of its performance, especially in lower FAR areas. On the other hand, LDA and EVM have similarly good detection and identification rates. Here, both algorithms can make use of information about the identities during training. For example, LDA computes its projection matrix so that all known unknowns are clustered together, and are far from any known subject. In opposition, EVM does not cluster the known unknowns, but only uses distances to them during training.

And precisely because EVM does not model the known unknowns, it is able to maintain its high performance when confronted with unknown unknowns from probe set O_2. On the other hand, LDA's performance dropped dramatically in Fig. 3(d) for low FAR values. We assume that the unknown unknowns do not cluster well in the LDA subspace and are, hence, more similar to gallery subjects. For larger FAR values, however, LDA still outperformed EVM. We attribute

this to the fact that LDA works well for a biased protocol [22], i.e., where identities in the training and test sets are shared. We assume that EVM, in opposition, is not favored by a biased protocol – as EVM does not use identity information of other subjects, but treats all distances to other subjects' features identically.

Another interesting point is that the discrepancy between the two modeling approaches, i.e., computing the maximum probability over three points and averaging the model features has an influence on the performance of EVM. While both in the closed-set evaluations and in the open-set evaluation with known unknowns, the average approach works better, it is the opposite in the open-set evaluations with unknown unknowns. It seems that with identities not seen during training, having more complex models results in a higher robustness with respect to rejecting unknown unknowns.

5. Discussion

Due to the fact that the number of unknown probe files is relatively low: $|\mathbf{K}| = 1334$ and $|\mathbf{U}| = 4069$, computing low false alarm rates, i.e, FAR < 0.001 was often not possible, and results in that range might not be statistically meaningful. Hence, the advantage of EVM over LDA in Fig. 3(d) might not be as significant as it looks. However, the advantage of EVM over raw cosine distances is obvious, both in the closed set ROC curve in Fig. 3(b), as well as in

the open-set evaluation in Fig. 3(d) since EVM performs better for almost any FAR value.

It is well-known [22] that LDA-based face recognition algorithms are highly favored by biased protocols like the one that we have introduced. This is due to the fact that LDA can make use of class information by clustering these classes together during training. For unbiased protocols Günther *et al.* [13] have found that LDA does not improve over simple distance computations in PCA subspace. Hence, LDA results on more realistic, i.e., unbiased datasets will most probably be lower.

In opposition, EVM does not make use of the classes during training. For each feature vector in a gallery template, only distances to all other subjects' feature vectors are computed to model the probability of inclusion. Theoretically, there is little difference whether these features belong to known or unknown subjects. Hence, we assume that it does not matter whether to query with subjects seen during training, or with unknown subjects, but we leave the verification of this assumption to future work. Based on this assumption, we can claim that EVM can handle unknown images better than LDA, and clearly better than using a simple thresholded cosine distance. Hence, we conclude that open-set recognition is better handled by modeling probability of inclusion with respect to gallery support and then thresholding on the posterior probability estimate, as opposed to thresholding raw similarities. Note that EVMs are also not limited to using simple distance functions between raw features. For example, a combination of LDA and EVM – first projecting the features into the LDA subspace and then applying an EVM in projected feature space – could be a viable approach.

Due to the relatively small size of the dataset, we did not split it up further into validation and test sets with mutually exclusive subjects. This is why we illustrated the performance of several hyperparameter choices on the test set, rather than use a subset of non-test data to select one set of hyperparameters. However, performance differed surprisingly little across hyperparameter choices (cf. Fig. 2), and choosing other parameters would not change our results.

6. Conclusion

In this paper we have shown that open-set face recognition is a difficult problem, and that simply thresholding similarity scores is a weak solution. We have experimented with two approaches that are often applied for face recognition: computing cosine distances on deep features, and applying linear discriminant analysis (LDA). Due to the biased nature of our evaluation protocol, LDA worked favorably over cosine in the open-set evaluations, but still performed poorly when tested with unknown unknowns. Hence, while LDA might be a proper choice for application when mainly known unknowns occur, in public areas (e.g.

in airports) with a high amount of passenger traffic, LDA will not be sufficient. Interestingly, LDA performed worst in the closed-set identification task, yet performed best in the verification task.

In order to model probabilities of inclusion with respect to gallery templates, we invoked the extreme value machine (EVM). Without making use of identity information during training, EVM was able to perform well in all of our tests, i.e., closed-set identification, verification and open-set identification. In all cases, EVM was able to beat the simple cosine distance, which demonstrates that modeling inclusion probabilities improves both closed and open-set identification as well as verification. Further, we assume that in an unbiased dataset, where training and test sets contain different identities, LDA will perform poorly while EVM will approximately maintain its performance. How well EVM performs with respect to other score normalization techniques such as Z-norm is left for future work.

Anyways, at a false alarm rate of 0.01 (meaning that 1 out of 100 unknown subjects are assigned to one subject in the gallery) only around 60 % of the gallery subjects were correctly identified by EVM. Revisiting our example in Sec. 1, a surveillance system in an airport that captures 100 persons per minute and queries each against a criminal database will have one false alarm per minute – which usually requires human interaction to resolve – while failing to identify 40 % of the criminals. Though this is a simplified example, it illustrates that open-set face identification is far from being solved, and additional research is required for real-time surveillance applications. While the open-set face identification protocol that we have introduced in this paper is a good start, research in the open-set identification space would benefit from larger databases that can be split into training, validation and testing sets, yet contain sufficiently many unknown unknowns to be able to calculate meaningful detection and identification rates at reasonable false alarm rates.

Acknowledgment

References

[1] A. Anjos, L. El Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel. Bob: a free signal processing and machine learning toolbox for researchers. In *ACM Conference on Multimedia Systems (ACMMM)*, 2012. 5

[2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1), 2000. 3

[3] A. Bendale and T. E. Boult. Towards open set deep networks. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. 3

[4] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *Transactions on Information Forensics and Security (TIFS)*, 9(12), 2014. 2, 3, 6

[5] C. M. Bishop. Pattern recognition. *Machine Learning*, 128, 2006. 3

[6] J.-C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep CNN features. In *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016. 2, 3, 4, 6

[7] J. S. del Rio, D. Moctezuma, C. Conde, I. M. de Diego, and E. Cabello. Automated border control e-gates and facial recognition systems. *Computers & Security*, 62, 2016. 1

[8] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *International Conference on Spoken Language Processing (ISCLP)*, 1998. 7

[9] H. K. Ekenel, L. Szasz-Toth, and R. Stiefelhagen. Open-set face recognition-based visitor interface system. In *International Conference on Computer Vision Systems (ICVS)*. Springer, 2009. 2, 3

[10] R. A. Fisher and L. H. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24. Cambridge University Press, 1928. 5

[11] P. Grother, R. J. Micheals, and P. J. Phillips. Face recognition vendor test 2002 performance metrics. In *International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*. Springer, 2003. 2

[12] M. Günther, L. El Shafey, and S. Marcel. *Face Recognition Across the Imaging Spectrum*, chapter Face Recognition in Challenging Environments: An Experimental and Reproducible Research Survey. Springer, 1st edition, 2016. 1, 4

[13] M. Günther, L. El Shafey, and S. Marcel. 2D face recognition: An experimental and reproducible research survey. Technical Report Idiap-RR-13-2017, Idiap, 2017. 8

[14] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015. 1

[15] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 1, 2, 3

[16] S. Huq, B. Abidi, S. G. Kong, and M. Abidi. *3D Imaging for Safety and Security*, chapter A Survey on 3D Modeling of Human Faces for Face Recognition. Springer, 2007. 1

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia (ACMMM)*, 2014. 5

[18] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015. 1, 3

[19] Y. LeCun, C. Cortes, and C. J. C. Burges. The MNIST database of handwritten digits, 1998. 3

[20] F. Li and H. Wechsler. Open set face recognition using transduction. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27(11), 2005. 2, 3

[21] S. Liao, Z. Lei, D. Yi, and S. Z. Li. A benchmark study of large-scale unconstrained face recognition. In *International Joint Conference on Biometrics (IJCB)*. IEEE, 2014. 2, 3

[22] Y. M. Lui, D. S. Bolme, P. J. Phillips, J. R. Beveridge, and B. A. Draper. Preliminary studies on the good, the bad, and the ugly face recognition challenge problem. In *Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, 2012. 7, 8

[23] A. J. O'Toole, P. J. Phillips, F. Jiang, J. Ayyad, N. Penard, and H. Abdi. Face recognition algorithms surpass humans matching faces over changes in illumination. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29, 2007. 1

[24] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference (BMVC)*, 2015. 1, 2, 3, 5

[25] P. J. Phillips, P. Grother, and R. Micheals. *Handbook of Face Recognition*, chapter Evaluation Methods in Face Recognition. Springer, 2nd edition, 2011. 2, 4

[26] E. M. Rudd, L. P. Jain, W. J. Scheirer, and T. E. Boult. The extreme value machine. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. to appear. 2, 3, 5

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015. 3

[28] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa. Triplet probabilistic embedding for face verification and clustering. In *Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2016. 3

[29] W. J. Scheirer, A. de Rezende Rocha, R. Michaels, and T. E. Boult. Meta-recognition: The theory and practice of recognition score analysis. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33, 2011. 5

[30] W. J. Scheirer, A. de Rezende Rocha, R. Micheals, and T. E. Boult. Robust fusion: Extreme value theory for recognition score normalization. In *European Conference on Computer Vision (ECCV)*. Springer, 2010. 3

[31] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward open set recognition. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(7), 2013. 3

[32] W. J. Scheirer, L. P. Jain, and T. E. Boult. Probability models for open set recognition. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(11), 2014. 3

[33] J. Stallkamp, H. K. Ekenel, and R. Stiefelhagen. Video-based face recognition on real-world data. In *International Conference on Computer Vision (ICCV)*. IEEE, 2007. 2, 3

[34] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Neural Information Processing Systems (NIPS)*, 2014. 2, 3

[35] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015. 2

[36] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014. 1, 3

[37] M. Tistarelli and Y. Sun. Cohort-based score normalization. In *Encyclopedia of Biometrics*. 2nd edition, 2015. 3

[38] W. S. Yambor. Analysis of PCA-based and Fisher discriminant-based image recognition algorithms. Master's thesis, Colorado State University, 2000. 5

[39] W. Zhao, A. Krishnaswamy, R. Chellappa, D. L. Swets, and J. Weng. *Face Recognition: From Theory to Applications*, chapter Discriminant Analysis of Principal Components for Face Recognition. Springer, 1998. 2, 3, 5