2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Extended YouTube Faces: a Dataset for Heterogeneous Open-Set Face Identification

Claudio Ferrari, Stefano Berretti, Alberto Del Bimbo
Media Integration and Communication Center (MICC)
Department of Information Engineering, University of Florence, Italy
{claudio.ferrari,stefano.berretti,alberto.delbimbo}@unifi.it

***Abstract*—In this paper, we propose an extension of the famous YouTube Faces (YTF) dataset. In the YTF dataset, the goal was to state whether two videos contained the same subject or not (video-based face verification). We enrich YTF with still images and an identification protocol. In the classic face identification, given a probe image (or video), the correct identity has to be retrieved among the gallery ones; the main peculiarity of such protocol is that each probe identity has a correspondent in the gallery (*closed-set*). To resemble a realistic and practical scenario, we devised a protocol in which probe identities are not guaranteed to be in the gallery (*open-set*). Compared to a closed-set identification, the latter is definitely more challenging in as much as the system needs firstly to reject impostors (*i.e.*, probe identities missing from the gallery), and subsequently, if the probe is accepted as genuine, retrieve the correct identity. In our case, the probe set is composed of full-length videos from the original dataset, while the gallery is composed of templates, *i.e.*, sets of still images. To collect the images, an automatic application was developed. The main motivations behind this work can be found in both the lack of open-set identification protocols defined in the literature and the undeniable complexity of such. We also argued that extending an existing and widely used dataset could make its distribution easier and that data heterogeneity would make the problem even more challenging and realistic. We named the dataset Extended YTF (E-YTF). Finally, we report baseline recognition results using two well known DCNN architectures.**[1]

I. Introduction

Since the pioneering work of Bledsoe et al. [1] in the mid '60s, for decades most of the research on face recognition focused on the definition of hand-crafted features (also referred to as "shallow" features) capable of capturing the traits of the face that best discriminate one subject from the others. For many years, these methods have been experimented on images acquired in cooperative contexts (indoor laboratories in most of the cases), with controlled conditions and a quite limited variability in terms of number of different identities, pose and illumination changes, image resolution, and so on. However, solutions based on classical learning methods and shallow features showed to be still quite not ready to cope with the large variability that occurs in the reality.

A large consensus has been now reached in the research community that Deep Convolutional Neural-Networks (DCNNs) can provide the right tools to perform face recognition in real and challenging conditions. Indeed, breakthrough results have been obtained using such technology on most of the existing benchmark datasets [2], [3], [4], [5], [6], [7]. Though CNNs were known since mid '80s, their effective deployment in real application contexts has been not possible till massive computation infrastructures and large quantities of data were available for training. In fact, one substantial innovation of DCNNs is the idea of letting the deep architecture to automatically discover low-level and high-level representations from labeled and/or unlabeled training data, which can then be used for detecting and/or classifying the underlying patterns. So, data are assuming an ever increasing relevance both for training and test. In particular, we observe three main directions to improve the effectiveness of training, and to increase the difficulty of testing: *(i)* large scale, *(ii)* large variability, and *(iii)* more heterogeneity (*i.e.*, mixed media).

For training, the importance of growing to a scale of million images is quite manifest in many works [2], [4], with the largest number of training images used by Google that in [5] trained on 200M photos of 8M people. For testing, the performance saturation on several common face recognition datasets makes it evident the need for more challenging benchmarks and protocols. Indeed, increasing the scale at which face recognition is performed, represents a natural way to make the recognition problem more challenging. In fact, several works demonstrated that the effectiveness of recognition substantially decreases when the number of gallery subjects does increase [8]. Large variability of the data is a second direction that is now pushed in the existing benchmarks (both for training and testing). In most of the cases, this is obtained by collecting real world images that naturally gather a large spectrum of variability, rather than using posed datasets. The goal is to increase the challenge of recognition by including more variability in terms of occlusion, illumination, expression, resolution, number of ethnicities, and so on. The shift from cooperative to "in the wild" datasets acquired without any subject cooperation [9] is a clear example of this trend. Mixing different media is the third direction followed to create more challenging face benchmarks. The idea here is that recognizing faces by comparing images vs. videos, and vice versa, is more difficult than performing face recognition based on still images or videos alone. One motivation for this is that videos have typically lower resolution than images, and naturally collect moving people that thus show more variability in their appearance. Another reason for proposing and evaluating mixed-media scenarios is that they correspond

[1]The dataset, metadata and protocols are available at https://www.micc.unifi.it/resources/datasets/e-ytf/

to the common case where images are included in the gallery, while videos are captured by monitoring cameras.

In addition to the points reported above, a further aspect which is gaining interest in face recognition is that of evaluating new protocols that go beyond the *verification* and *identification* ones. A division that is making its way in face identification is that between *closed-* and *open-set*. The former assumes that all probe images used in the evaluation contain identities of subjects that are enrolled in the gallery. But real systems cannot make this closed-set assumption, since only a fraction of probe identities are in the gallery. Instead, the *open-set* protocol assumes face identification to be able to reject/ignore those probes that correspond to unknown identities. This latter protocol is considered as much more difficult than the closed-set one, and its use for evaluating face recognition methods is still at the beginning. The "open world" modalities are also coming into the scene. In these protocols, the probe identities that are unknown should be automatically enrolled into the gallery (auto-enrollment).

In this paper, we propose a new dataset, that we call Extended-YTF (E-YTF), for training and testing face recognition algorithms, which extends the well-known and largely used YouTube Faces (YTF) benchmark. With respect to the points addressed above, the extension we propose here contributes mainly on the data heterogeneity and the protocols. First, we extend YTF–that only includes videos–with still images, thus enabling a mixed-media face recognition modality. To this end, we devised a semi-automatic image gathering tool that collects, from the Internet, still images with the same identity of the celebrities in YTF. These images have been taken in the wild and include large variability in terms of expression, illumination, and occlusion. The pose of the face has variations too, though extreme cases of full side views occur rarely. Secondly, while face verification from videos is the only protocol used in YTF, in the proposed E-YTF we add a closed- and an open-set identification protocol for mixed-media. In the *heterogeneous and closed-set* case, images are used as gallery and videos of only the same subjects of the gallery are used as probes; In the *heterogeneous and open-set* case, the videos used as probes also include a large number of identities that are not present in the gallery. These two new protocols are both challenging and have not been considered much in the literature so far. The proposed protocol is completed by the definition of appropriate splits for training and testing. Finally, using these data and protocols, we also present baseline results obtained with two well-known DCNN architectures, namely, AlexNet [10] and VggFace [4].

The rest of the paper is organized as follows: In Sect. II, we summarize the literature on recent benchmarks for face recognition and the related protocols; In Sect. III, the way we collected the data is expounded; The protocols that we propose and evaluate on the newly created E-YTF dataset are described in Sect. IV; Baseline results for E-YTF are reported in Sect. V using the defined protocols and two CNN architectures; Finally, we draw conclusions in Sect. VI.

II. Related Work

The literature relevant to this work mostly concerns with the existing benchmarking datasets for face recognition and the related evaluation protocols.

The Labeled Faces in the Wild (LFW) dataset as proposed by Huang et al. [9], was the first largely used benchmark that included images acquired in unconstrained domains. The database contains 13,233 target face images of 5,749 different individuals. Of these, 1,680 people have two or more images in the database, while the remaining 4,069 people have just a single image. All the images are the result of the Viola-Jones face detector, subsequently rescaled and cropped to a fixed size. False positive face detections were manually eliminated, along with images for whom the name of the individual could not be identified. Sengupta et al. [11] proposed the Celebrities in Frontal-Profile (CFP) dataset with the intent of isolating the factor of pose variation, in terms of extreme poses like profile, along with other "in the wild" variations. The dataset contains 10 frontal and 4 profile images, where many features are occluded, of 500 individuals. Similar to LFW, 10 splits are defined, each containing 350 same and 350 not-same pairs. The task is frontal to profile face verification "in the wild". Both LFW and CFP include images only. Wolf et al. [12], instead, proposed the YouTube Faces (YTF) dataset that collects videos from YouTube and it is specifically designed to study the problem of video based face verification. The dataset contains 3,425 videos of 1,595 subjects, and the task is to decide whether two video sequences contain the same subject or not.

The above datasets have two main limitations. On the one hand, the number of different identities is relatively small, with a scale of thousands. On the other, each dataset is targeted to a specific media (either images or videos). Thus, other datasets have been proposed in the literature that aim to go one step further with respect to such limitations. Motivated by the performance saturation on some major benchmarks (*i.e.*, LFW), where a number of algorithms achieved near to perfect score, surpassing human recognition rates, Kemelmacher-Shlizerman et al. [8] proposed evaluations at the million scale by assembling the MegaFace dataset. This dataset includes 1M photos that capture more than 690K different individuals. The related challenge evaluates performance of algorithms with increasing number of "distractors" (going from 10 to 1M) in the gallery set. Both identification and verification protocols were proposed, using two sets as probes: The FaceScrub dataset [13], which includes 100K photos of 530 celebrities; and the FG-NET aging dataset [14], [15], which includes 975 photos of 82 people. A face image dataset at one million scale, called MS-Celeb-1M, was also proposed by Guo et al. [16]. As benchmark task they defined the recognition of 1M celebrities from their face images, by using all the possibly collected face images of these individuals on the web as training data.

Despite the importance of rigorous testing data for evaluating face recognition algorithms, all major publicly available faces-in-the-wild datasets are constrained by the use of a com-

modity face detector, which limits, among other conditions, pose, occlusion, expression, and illumination variations. To mitigate these constraints, Klare et al. [17] released the NIST IJB-A dataset. It is a publicly available media in the wild dataset containing 500 subjects with manually localized face images. Key features of the IJB-A dataset are: *(i)* full pose variation, *(ii)* joint use for face recognition and face detection benchmarking, *(iii)* a mix of images and videos, *(iv)* wider geographic variation of subjects, *(v)* protocols supporting both closed-set identification with distractors (1:N search) and verification (1:1 comparison), *(vi)* an optional protocol that allows modeling of gallery subjects, and *(vii)* ground truth eye and nose locations. The dataset has been developed using 1,501,267 million crowd sourced annotations. In summary, this dataset presents more challenging "in the wild" face acquisitions (*e.g.*, full pose variations), for mixed media. The relatively low number of impostor and genuine matches per split in the IJB-A protocol limits the evaluation of an algorithm at operationally relevant assessment points. Building upon IJB-A, the work by Whitelam et al. [18] introduced the IARPA Janus Benchmark-B (NIST IJB-B) dataset, a superset of IJB-A. It consists of 1,845 subjects with human-labeled ground truth face bounding boxes, eye/nose locations, and covariate metadata such as occlusion, facial hair, and skin tone for 21,798 still images and 55,026 frames from 7,011 videos. IJB-B was also designed to have a more uniform geographic distribution of subjects than that of IJB-A. Test protocols for IJB-B represent operational use cases including access point identification, forensic quality media searches, surveillance video searches, and clustering. Summarizing, this dataset allows also open-set identification from mixed media.

Much research has been conducted on both face identification 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 is 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 corresponding to unknown identities. Thus, there is now interest in shifting face recognition benchmarks toward more difficult scenarios. The open-set protocol has been evaluated in a recent face recognition challenge by Günther et al. [19], which addresses the next step in the direction of automatic detection and identification of people from outdoor surveillance cameras. Results show that open-set face recognition is currently weak and requires much more attention. One recent proposal in this direction has been presented by Günther et al. [20]. They started from the observation that a widespread approach to the open-set identification problem is that of thresholding verification-like scores. They evaluated the goodness of this solution by first formulating an open-set face identification protocol based on the canonical LFW dataset, where additionally to the known identities, they introduced the concepts of *known unknowns* (known, but uninteresting persons) and *unknown unknowns* (people never seen before). Then, they compared three algorithms for assessing similarity in a deep feature space under an open-set protocol: thresholded verification like scores, linear discriminant analysis scores, and an extreme value machine (EVM) probabilities. Results suggest that thresholding EVM probabilities, which are open-set by design, outperforms thresholding verification like scores.

Fig. 1. Example images of 4 subjects from the YTF dataset taken from the web (top row); despite the uncontrolled conditions, the good quality and the cooperation of the subjects can be appreciated. Imagery of the same individuals from the original dataset (bottom row).

III. Data collection

The YTF dataset is composed of video sequences only that will be used as probe set in the proposed identification protocol. The gallery is instead built collecting new data of the same individuals in the form of still images, rather than using part of the available sequences. In realistic applications, gallery sets are usually made up of good quality images, which are often taken enrolling the subjects in constrained conditions. On the other hand, the probe query may come in whatever form and quality, *e.g.*, video surveillance sequences. This motivated us to choose the video sequences of the original dataset as probe, while the gallery images are collected from the web. As shown in Fig. 1 (top row), celebrities imagery are mostly taken professionally, and despite the unconstrained conditions, they often result largely frontal and cooperative; this reasonably approximates a real scenario. In so doing, the problem of matching data from heterogeneous sources is also explored. In fact, as shown in [21], such diversity negatively affects the performance of systems based on DCNNs.

A. Collection Procedure

To gather a set of images for each identity, a web collector has been implemented. Such application takes a list of names, queries 3 different search engines and downloads a user-defined number of results. A face detector is then run over each downloaded image and the cropped faces are stored. After that, an automatic procedure to filter out possibly wrong identities or detections is run. This is necessary since *(i)* it is rather likely that each downloaded image contains more than one subject and *(ii)* web searches can return photos of persons that are not the subject of interest but someone that is somehow related to him. With this aim, for each subject, we use the pre-trained

TABLE I
MAIN FEATURES OF FACE BENCHMARK DATASETS. E-YTF IS THE EXTENDED YTF DATASET PROPOSED IN THIS WORK

Dataset	#subjects	#images	#videos	train partition	pose variation	protocol	open-set	mixed media
LFW [9]	5,749	13,233	0	yes	moderate	verification	-	no
CFP [11]	500	7000	0	yes	frontal / profile	verification	-	no
MegaFace [8]	690K	1M	0	no	full	identification / verification	no	no
MS-Celeb-1M [16]	100K	10M	0	yes	moderate	retrieval	no	no
IJB-A [17]	500	5,712	2,085	yes	full	identification / verification	no	yes
IJB-B [18]	1,845	21,798	7,011	no	full	identification / verification	yes	yes
YTF [12]	1,595	0	3,425	yes	moderate	verification	-	no
E-YTF	**1,595**	**38,667**	**3,425**	**yes**	**moderate**	**identification**	**yes**	**yes**

VggFace model [4] to extract a face descriptor from all its collected images. Subsequently, we select a random subset of the descriptors to train a binary SVM classifier; all the images that are not classified as belonging to the subject are marked as wrong. This procedure is highly effective if the most of the imagery actually contains the correct subject. In this scenario, outliers are rejected pretty accurately. In more unfortunate cases, where the number of correct images is small like homonymy cases or large amount of group pictures, the automatic filtering can fail. In fact, choosing the right images with which train the classifier is non-trivial. To refine the latter process, we developed a web application that can be used to check the result of the automated step and, in case, correct the labeling. The application shows, for each identity, all its images and allows the user to mark the wrong ones. The images that have been previously marked by the automated procedure are shown. The images are released along with the bounding box annotations, which were obtained using the TinyFaces face detector [22].

B. Statistics

In this section, we report some statistics of the newly collected images; 677 identities have been retained, for a total of $38,677$ still images. This is the final number resulting after the filtering procedure. Identities from the original dataset which were less well-known, resulted in uncertain web searches and most of the downloaded images associated to such individuals were wrong. Motivated by the goal of building an open-set protocol, we decided to discard the identities which were considered to contain too many errors. The average number of images per identity is 57, with a minimum of 1 image and a maximum of 130 images. The distribution of images per subject is shown in Fig. 2. The average video sequence length is instead 180 frames, with a minimum of 48 frames and a maximum of $2,157$. Table I summarizes the characteristics of several face benchmarks in comparison to the E-YTF proposed in this paper.

IV. PROTOCOLS AND PERFORMANCE METRICS

In the following, the two protocols, *i.e.*, open-set and closed-set identification, and the metrics used for evaluation of the proposed E-YTF are described.

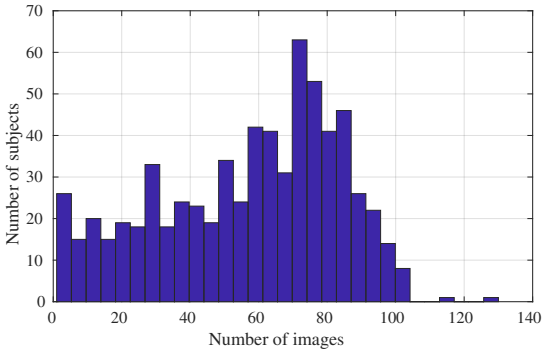
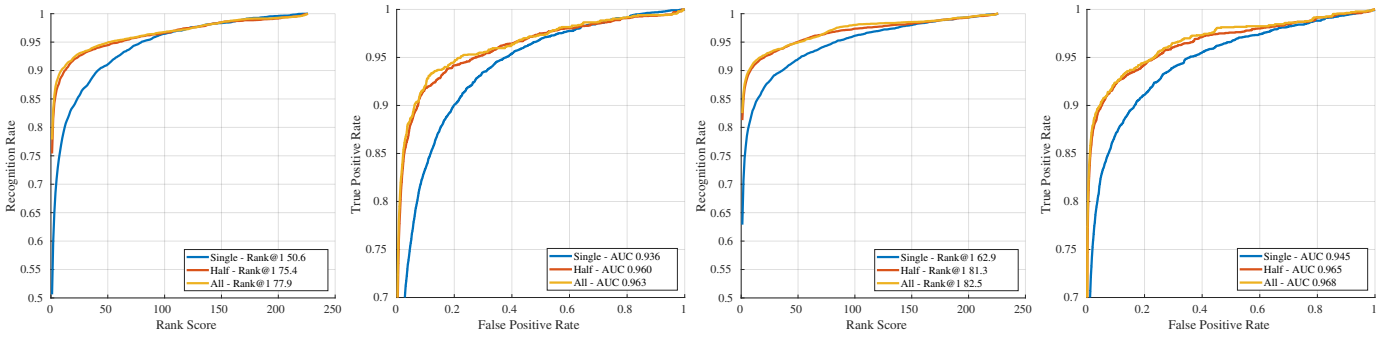


Fig. 2. Number of identities in function of the number of images.

A. Protocols

The evaluation is to be carried out on 10 splits. For each split, we divide the data into train and test set randomly shuffling the identities, *i.e.*, identities in the train set do not appear in the test set. Following previous works [17], $2/3$ of the identities have been included in the train set, while the remaining $1/3$ constitutes the test set. The train set contains both still images and video frames, while the test set is in turn divided in a probe set and 3 gallery sets; the latter contain templates of still images (one template per identity), which are defined based on the number of images used to build the template: *(i) Single*: templates of single images, which are selected randomly; *(ii) Half*: templates of half of the total images of the subject, chosen randomly; *(iii) All*: all the available images of the subject are used to build the templates. This aspect is relevant to deepen the impact of differently sized templates in the matching. The probe set is instead made up of the video sequences. The search is conducted at video-level, *i.e.*, the decision must be taken considering the whole sequence.

The latter setup is used both for the closed-set and the open-set protocols. In the closed-set, the probe identities coincide with the gallery ones; in the open-set, all the identities of the original dataset are used. In this way, some probe subjects do not have a mate in the gallery. Note that in the open-set protocol, the additional probe identities are also disjoint from the training set. In the closed-set, for each split $\sim$ 100K video frames are used to search into the galleries, which include $\sim$ 12K, $\sim$ 6K and 226 images, respectively, for the *All*,



(a) (b) (c) (d)

Fig. 3. Average CMC and ROC across 10 splits for AlexNet (a-b) and VggFace (c-d), and for the different templates size. The *min* distance was used.

TABLE II
EVALUATION SETUPS OF IDENTIFICATION BENCHMARK DATASETS

Dataset	#ID	#train-ID	evaluation	#gallery-ID	#probe-ID
MegaFace [8]	690K	–	2 probe sets	690K	530 / 975
IJB-A [17]	500	333	10 splits	167	167
IJB-B [18]	1,845	–	2 gallery sets	931 / 914	1.845
E-YTF (closed)	**677**	**450**	**10 splits**	**227**	**227**
E-YTF (open)	**1595**	**450**	**10 splits**	**227**	**1145**

TABLE III
RANK@1 AND AUC SCORES IN FUNCTION OF THE TEMPLATE SIZE AND THE DIFFERENT DISTANCE MEASURES

	Rank@1				AUC			
	Half		All		Half		All	
Net	min	min+mean	min	min+mean	min	min+mean	min	min+mean
VggFace	81.3	**83.3**	82.5	**84.2**	96.5	**96.9**	96.8	**97.1**
AlexNet	75.4	**80.4**	77.9	**82.1**	96.0	**96.5**	96.3	**96.7**

Half and *Single* cases. In the open-set, instead, the probe frames are ~ 450K. In Table II the evaluation setups of some identification benchmark datasets are reported in comparison with the proposed E-YTF.

B. Metrics

For the closed-set identification protocol, we employ two measures: the Receiver Operating Characteristic (ROC) curve, and the Cumulative Match Characteristic (CMC) curve. The first measures the trade-off between sensitivity and specificity, while the second measures the ranking accuracy; the latter calculates the percentage of probe searches that return a true positive within the first k sorted ranks. The open-set protocol, as in [18], is instead evaluated in terms of the Identification Error Trade-off (IET), which shows how the false positive rate (FPR) varies with respect to the false negative rate (FNR). FPR is the proportion of searches that return at least one incorrect gallery candidate above a threshold t, while FNR is the proportion of searches that did not return any correct gallery candidate above the same threshold. These metrics give a fairly complete overview of the performance of a recognition algorithm.

V. BASELINE RESULTS

We report some baseline results obtained with a standard face recognition pipeline, that is composed of the following main steps: *(i)* detection, *(ii)* alignment, *(iii)* representation, *i.e.*, feature extraction, and *(iv)* matching.

A. Pipeline

Both the still images and the video sequences come along with bounding box annotations, thus the detection step was skipped. Following the guidelines in [21], the provided bounding boxes were enlarged so as to include the whole head and the alignment step was bypassed also. The face crops and their horizontally flipped version were then fed to two different pre-trained CNN architectures to extract feature descriptors; the final descriptor is obtained as the average of the two. We employed the publicly available VggFace model [4] and the AlexNet architecture [10], trained as in [21]. For each video sequence in the probe set, we computed the average descriptor from all the frames. The motivation for this is two-fold: first, the YTF video sequences are rather short and thus the variability in the appearance is supposed to be limited; in this sense, it also helps in attenuating the effect of outliers. Secondly, it allows a much faster matching procedure.

Finally, we employed the cosine distance to match probe and gallery. Being the gallery composed of templates, one needs to derive a single scalar value from all the distances computed between the video sequences and each image in the templates. We followed two strategies to achieve this: one solution employs a simple nearest-neighbor approach, in which the minimum of the distances is used as final measure (*min*). The second solution sums up the minimum and the average distance (*min+mean*). The latter grounds on the idea that, similarly to computing the average descriptor, the average distance can help in reducing the impact of possible outliers.

B. Closed- and Open-set Identification

Baseline results for the closed-set protocol are reported in Fig. 3; the plots show results obtained using the *min* distance as final measure, for the different gallery templates size. The outcomes show that there is a clear advantage in having gallery templates composed of more than one image, mostly in terms of Rank@1 accuracy. On the other hand, the small difference

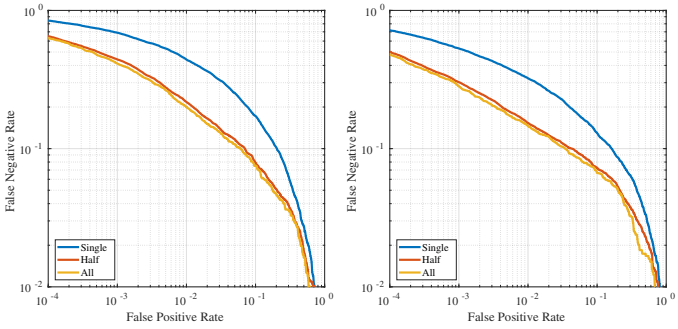


Fig. 4. Average IET performance across 10 splits for AlexNet (a) and VggFace (b), and for the different templates size. The *min+mean* distance was used.

between the "Half" and "All" cases suggests that the quality of the images in the templates is likely to be more important than the template size itself. However, choosing a suitable matching strategy can radically change the scenario. Table III shows a comparison between the two final distance measures; despite being relatively more complex, the use of the *min+mean* distance actually improves upon the simple nearest-neighbor, more significantly for the AlexNet model.

Results for the open-set protocol are reported in Fig. 4. Here, the final measure used was the *min+mean* distance, since better performance were achieved in the closed-set protocol.

VI. Conclusion

In this paper, we proposed the Extended YouTube Faces (E-YTF) dataset, which is an extension of the widely famous YouTube Faces (YTF). Our proposed extension enlarges the original dataset with still images collected from the web for a subset of the identities in the original dataset ($38,667$ new images in total). Along with the additional data, we devised two identification protocols: *closed-set* and *open-set*. The main peculiarity of open-set identification is that probe identities may not have a correspondent in the gallery. The latter is lately gaining increasing attraction in the Computer Vision field, mostly because of its intrinsic difficulty and applicability in the real world. Also, results on standard benchmark datasets are relentlessly saturating, allowing the community to move its focus towards more realistic and practical scenarios. We believe that extending an already popular dataset and considering standard and widely used performance measures can help in this transition and facilitate the comparison of methods developed to this end. Baseline results obtained with a standard recognition pipeline based on state-of-the-art Convolutional Neural Networks are finally reported.

Acknowledgment

The authors would like to thank Giuseppe Lisanti, Francesco Turchini, Andrea Salvi, Claudio Baecchi, Leonardo Galteri, Tiberio Uricchio and Federico Becattini for their precious help in annotating and refining the dataset.

References

[1] W. W. Bledsoe, "Some results on multicategory pattern recognition," *Journal of the ACM*, vol. 13, no. 2, pp. 304–316, 1966.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[3] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 1891–1898.

[4] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conf. (BMVC)*, vol. 1, no. 3, 2015, p. 6.

[5] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 815–823.

[6] X. Yin and X. Liu, "Multi-task convolutional neural network for pose-invariant face recognition," *IEEE Trans. on Image Processing*, vol. 27, no. 2, pp. 964–975, Feb 2018.

[7] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer *et al.*, "The challenge of face recognition from digital point-and-shoot cameras," in *IEEE Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, 2013, pp. 1–8.

[8] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 4873–4882.

[9] 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, Oct. 2007.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Int. Conf. on Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[11] S. Sengupta, J. C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *IEEE Winter Conf. on Applications of Computer Vision*, 2016, pp. 1–9.

[12] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 529–534.

[13] H. W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *IEEE Int. Conf. on Image Processing*, 2014, pp. 343–347.

[14] I. Kemelmacher-Shlizerman, S. Suwajanakorn, and S. M. Seitz, "Illumination-aware age progression," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 3334–3341.

[15] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes, "Overview of research on facial ageing using the fg-net ageing database," *IET Biometrics*, vol. 5, no. 2, pp. 37–46, 2016.

[16] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large scale face recognition," in *European Conf. on Computer Vision*. Springer, 2016.

[17] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1931–1939.

[18] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother, "Iarpa janus benchmark-b face dataset," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 592–600.

[19] M. Günther, P. Hu, C. Herrmann, C. Chan, M. Jiang, S. Yang, A. R. Dhamija, D. Ramanan, J. Beyerer, J. Kittler, M. A. Jazaery, M. I. Nouyed, C. Stankiewicz, and T. E. Boult, "Unconstrained face detection and open-set face recognition challenge," *CoRR*, vol. abs/1708.02337, 2017. [Online]. Available: http://arxiv.org/abs/1708.02337

[20] M. Günther, S. Cruz, E. M. Rudd, and T. E. Boult, "Toward open-set face recognition," *CoRR*, vol. abs/1705.01567, 2017. [Online]. Available: http://arxiv.org/abs/1705.01567

[21] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo, "Investigating nuisance factors in face recognition with dcnn representation," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 583–591.

[22] P. Hu and D. Ramanan, "Finding tiny faces," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 951–959.