Toward More Realistic Face Recognition Evaluation Protocols for the YouTube Faces Database

Yoanna Martínez-Díaz, Heydi Méndez-Vázquez, Leyanis López-Avila
Advanced Technologies Application Center (CENATAV)
7A ♯21406 Siboney, Playa, P.C. 12200, Havana, Cuba
{ymartinez,hmendez,lelopez}@cenatav.co.cu

Leonardo Chang
Tecnológico de Monterrey,
Estado de Mexico, Mexico
lchang@itesm.mx

L. Enrique Sucar
INAOE,
Puebla, Mexico
esucar@inaoep.mx

Massimo Tistarelli
University of Sassari,
Sassari, Italy
tista@uniss.it

Abstract

One of the key factors to measure the progress of a research problem is the design of appropriate evaluation protocols defined on suitable databases. Recently, the introduction of comprehensive databases and benchmarks of face videos has had a great impact on the development of new face recognition techniques. However, most of the protocols provided for these datasets are limited and do not capture requirements of unconstrained scenarios. That is why sometimes the performance of face recognition methods on current benchmarks seems to be saturated. To address this lack, the tendency is to collect new datasets, which is more expensive and sometimes the main the problem is not the data but the protocols. In this work, we propose new relevant evaluation protocols for the YouTube Faces database (REP-YTF) supporting face verification and open/closed-set identification. The proposal better fits realistic face recognition scenarios and allows us to test existing algorithms at relevant assessment points, under different openness values and taking both videos and images as the gallery. We provide an extensive experimental evaluation, by combining several well-established feature representations with three different metric learning algorithms. The obtained results show that by using the proposed evaluation protocols, there is room for improvement in the recognition performance on the YouTube Faces database.

1. Introduction

Despite face recognition in videos has seen significant breakthroughs in the last few years, it is still considered an unsolved problem in unconstrained scenarios, as supported by the poor performance of state-of-the-art methods on practical applications (e.g., video surveillance, access control, etc.). In order to measure the advances in this area and to compare existing algorithms, appropriate evaluation protocols defined on suitable databases are needed.

Several databases are available for the evaluation of unconstrained face recognition in videos [2, 6, 8, 9, 14, 18, 19]. However, few of them have been released with operationally relevant protocols. That is why sometimes the performance of face recognition methods on current benchmark datasets seems to be saturated achieving near to perfect score. As a result, the research community usually focuses on collecting new data when the problem is not mainly the data but the limitations of the existing protocols that do not exploit all the available data and do not capture the requirements of operationally unconstrained scenarios. For example, the limited ability to evaluate algorithms at operationally relevant False Acceptance Rate (FAR) (e.g., 0.01% and 0.1%).

The YouTube Faces (YTF) database [19] is a leading benchmark for video face recognition, containing videos from a large number of persons. Several state-of-the-art methods [4, 5, 13, 16] have been evaluated and compared on this database, achieving accuracies higher than 90%. However, the standard protocol of the YTF database only considers the face verification scenario with a limited number of matching comparisons (2,500 genuine and 2,500 impostor matches), which do not provide a complete picture on the capabilities of the methods that are tested on it. For example, when evaluating performance at FAR = 0.1% only less than three impostor matches are available, which is not statistically sound. Moreover, a classification accuracy of 97.3% may imply a FAR of 2.7%, which is not suitable for unconstrained face recognition applications in which usually low FAR values are demanded.

In order to address the above limitations, we design new relevant evaluation protocols for the YTF database (REP-YTF), which better fit operational face recognition systems. Specifically, we propose a new face verification protocol that makes full use of the whole database allowing the evaluation at low FAR values. Moreover, open/closed-set identification protocols are designed considering different gallery sizes, as well as, both video-to-video and video-to-image comparisons. Table 1 summarizes the main differences between the standard protocol and our proposal for face recognition evaluation in the YTF database.

We test several methods under the proposed REP-YTF, including well-known representations that have been previously evaluated by using the standard protocol of the YTF database, which are combined with different metric learning algorithms. The obtained results highlight that unconstrained video face recognition is still an unsolved and challenging problem, even in existing benchmark datasets such as YTF. The baseline code and evaluation scripts are available online [1].

The main contributions of the proposed REP-YTF are:

- It is clear and easy to understand.
- A more statistically sound evaluation at low FAR values is provided, thanks to the fact that a greater number of impostor comparisons is considered for face verification.
- Both open and closed-set face identification experiments are included using different gallery sizes.
- Besides video-to-video, video-to-image comparisons are included, thus a larger number of methods for different scenarios can be evaluated.
- The conducted experiments with 21 face recognition approaches (seven face representations with three metric learning algorithms) show a contrary perception of recognition performance on YTF being saturated.
- It is proved that more appropriate test protocols can be defined using existing benchmarks (e.g., YTF) to provide a more realistic evaluation of face recognition methods.
- It is publicly available to encourage and support algorithms development for unconstrained face recognition in videos.

The paper is organized as follows. Section 2 describes the YTF database benchmark as well as the limitations of its standard protocol. Section 3 details the proposed REP-YTF. The baseline methods used for the evaluation are presented in Section 4. Section 5 provides a large set of experiments illustrating the performance of the evaluated methods on the YTF database under the proposed protocols. Section 6 concludes the paper.

2. The YouTube Faces Database

The YouTube Faces database [19] is a large video dataset designed for unconstrained face verification in videos. It contains 3,425 videos of 1,595 subjects with significant variations in expression, illumination, pose, resolution, and background. An average of 2.15 videos are available for each subject. The average length of a video clip is 181.3 frames.

The standard protocol of the YTF database provides a pair-matching benchmark for a ten-fold cross-validation testing. Specifically, 5,000 video pairs from the database were randomly collected, half of which are pairs of videos of the same person and the other half of different subjects. These pairs were divided into ten splits, each one containing 250 ‘same’ and 250 ‘not-same’ pairs. These pairs were divided ensuring that, if videos of a subject appear in one split, no video of that subject is included in any other split. The test procedure consists of using the defined pairs of each split for algorithm evaluation and report the obtained performance. Hence, in the whole benchmark, although there are a large number of potential comparisons, only a reduced number of genuine and impostor matching scores are computed for classification.

Currently, the best-performing methods on the YTF database report more than 90% of accuracy by using deep learning-based representations [4, 5, 13, 16]. Nevertheless, this performance is too optimistic for most practical systems where the evaluation needs to be at low FARs like 0.01%, which is not possible to measure with the standard protocol since less than three impostor scores are available. Moreover, in the defined 5,000 video pairs, not all the videos from the database are included. There are more than 190 videos which are not taken into account. In addition, by considering only a face verification scenario, is not possible to provide all the capabilities of the face recognition methods.

Since that the standard YTF protocol is very limited and does not consider the whole available data, we exploit all the 3,425 face videos and propose new relevant evaluation protocols (REP-YTF) supporting face verification and open/closed-set face identification scenarios.

3. REP-YTF Description

In this section, we describe in detail the experimental setting of the designed protocols and the performance measures used for the evaluation.

3.1. Experimental Setting

We divided the YTF database into ten random trials of training and test sets. On each trial, we ensure that videos from subjects that are included in the training set are not considered in the test set. The training set of each trial in-

[1] http://www.cenatav.co.cu/doc/code/REP-YTF.zip

	Standard Protocol	REP-YTF
Used all available data	No	Yes
Closed-set identification protocol	No	Yes
Open-set identification protocol	No	Yes
Face verification protocol	Yes	Yes
# Genuine comparisons (not-duplicated)	2,500	2,277
# Impostor comparisons (not-duplicated)	2,500	3,314,989

Table 1: Differences between the standard protocol of the YTF database and the REP-YTF.

cludes the videos of 395 subjects from which 243 subjects on average have more than one face video. As a result, each training set contains more than 800 face videos available for the algorithms to build models and learn face variations. The testing sets are composed of the videos from the remaining 1,200 subjects, where about 744 subjects have at least one video, resulting in more than 2500 face videos tested on average for each trial.

For the verification protocol, the test set of each trial is used to compute the matching scores by face recognition algorithms for performance evaluation. On average, 2,277 genuine comparisons and 3,314,989 impostor comparisons are obtained in each trial, ensuring that they are not-duplicated. Thus, over 3,317,266 video pairs comparison scores are computed for each trial. This allows us to evaluate face recognition algorithms at low FARs. For example, at FAR = 0.1% there are more than 3,300 impostor comparisons available for each trial, which is more statistically sound.

In the case of identification, for each trial, we randomly partitioned the test set into three subsets, the gallery set G, the genuine probe set P_G, which contains the same subjects included in G but with different videos, and the impostor probe set P_I, which consists of subjects that are not present in G. In the closed-set identification protocol, P_G is evaluated against G, while in the open-set identification protocol, P_I is also tested. Besides, to evaluate the performance of the methods with different gallery sizes, this partitioning procedure is repeated three times, varying the openness (Op), that is defined as the ratio between the genuine comparisons and the impostor comparisons in the probe set. As a result, in each trial, three different configurations of the test set are obtained by using the openness values: 0.2, 0.5 and 0.9. In Table 2, all the described experimental setting is summarized.

In addition, we designed two kinds of galleries: one composed by a face video per subject and the other by a single image per subject. In the last case, to simulate real applications where the gallery image quality is usually good, we selected for each subject, the best face frame from the video gallery. This means that the frame with a frontal or near frontal pose, uniform lighting, neutral expression, no occlusion, and no blur is used.

			#Subjects	#Videos
Train			395	849
Test	Verification		1,200	2,576
	Op (0.2)	G	200	200
		P_G	200	370
		P_I	1,000	2,005
	Op (0.5)	G	400	400
		P_G	400	728
		P_I	800	1,448
	Op (0.9)	G	533	533
		P_G	533	975
		P_I	667	1,068

Table 2: Experimental setting for REP-YTF. Numbers are averaged over the 10 trials. G, P_G and P_I are the gallery, the genuine probe set and the impostor probe set, respectively.

3.2. Performance Metrics

In the face verification scenario, the goal is to decide whether two face videos belong or not to the same identity. In this case, we used the Receiver Operating Characteristic (ROC) curve and the Equal Error Rate (EER) as performance metrics. ROC curve is a plot of the True Acceptance Rate (TAR) versus the FAR by changing the decision thresholds, while EER is the point on the ROC curve where the errors of false acceptance and false rejection are the same.

Open-set identification consists of two steps: first, to determine if the identity of a face in the probe is present in the gallery or not; and, if it is, to find the top-k most similar faces in the gallery. Closed-set identification is a particular case of open-set identification, which assumes that the subject in the probe is present in the gallery. For evaluating open-set identification, detection and identification rate (DIR) [15] and FAR performance metrics are considered, while the Cumulative Match Characteristic (CMC) curve [15] is used for the case of closed-set identification.

Both for face verification and identification, the performance metrics are computed and averaged over the ten random trials, and the standard deviation is also reported.

4. Baselines for Video Face Recognition

As one of the most used benchmark on video face recognition, YTF database has largely promoted the research on feature descriptors, metric learning algorithms, and deep learning-based representations. Thus, we examine and study the performance of some well-established representations combined with several metric learning algorithms under the REP-YTF.

First of all, we used the Local Binary Patterns (LBP) descriptors provided with the YTF database[2]. These descriptors were extracted taking into account different pose based strategies as described in [19]. Specifically, we used the most frontal pose and the smallest head rotation angle strategies. In the first case, two videos will be compared based on the LBP descriptors of the frame of each video with the most frontal face pose and in the second one, the LBP descriptors of the frames with the smallest head rotation angle between them.

On the other hand, we evaluated three representations based on the Fisher Vector (FV) encoding of local descriptors. Specifically, we tested the Video Fisher Vector Faces (VF^2) descriptor[12] that encodes SIFT features, and the Binary Video Fisher Vector Faces (BinVF^2) [10] and the Logistic Binary Video Fisher Vector Faces (LBinVF^2) [11], which efficiently encode BRIEF descriptors. In all the cases, we used the parameters given by the authors in their corresponding papers.

Deep convolutional neural networks trained with massive labeled outside data have reported the best performance on the YTF database. We used the descriptors obtained from two pre-trained deep networks which have been applied on face recognition. First, the VGG-Face descriptors which are extracted by using the off-the-shelf CNN model based on the VGG-Very-Deep-16 CNN architecture described in [13]. The other is the model provided in Dlib open source machine learning library [7], which is a ResNet network with 27 convolutional layers, inspired on the ResNet-34 network from [5].

All the obtained descriptors are projected into a 200-dimensional PCA subspace in order to reduce their dimension and then, a metric learning algorithm is used to perform the comparison. The common objective of metric learning is to learn a suitable distance function to reduce the distance of positive pairs and enlarge the distance of negative pairs at the same time. Specifically, we tested the Large Margin Nearest Neighbor (LMNN) [17], the Joint Bayesian (JB) [3] and the Linear Discriminant Analysis [1].

In Large Margin Nearest Neighbor (LMNN) [17], the goal is to learn a Mahalanobis distance metric for kNN classification, where the distances are viewed as generalizations of Euclidean distances. This approach has the advantage of improving the original Euclidean distance from a classification perspective and, in some cases, to provide a lower-dimensional embedding of the data. The Joint Bayesian (JB) approach [3] models the joint distribution of feature vectors of a pair of face images belonging to the same or different subjects and uses the log-likelihood ratio of intra- and inter-class probabilities as the similarity measure. Linear Discriminant Analysis (LDA) [1] is a type of metric learning method which learns a projection such that it maximizes the inter-class scatter over the intra-class scatter. This is solved using a closed-form expression based on a generalized eigenproblem.

5. Experimental Evaluation

In this section, we analyze the results obtained by using the proposed evaluation protocols. Specifically, we test the baselines methods for video face recognition described previously in Section 3. In total, 21 face recognition approaches (7 face representations, each one with three different metric learning algorithms) are evaluated.

For the VGG-Face network, the input has to be a face image of size 224×224 pixels. For the remaining methods, all video frames were center cropped to 150×150 pixels.

5.1. Face Verification

Figures 1a, 1b and 1c show the verification ROC curves of the different face representations with LMNN, JB and LDA metric learning algorithms, respectively. Besides, in Table 3 the mean TAR at FAR values of 0.1% and 1%, and the mean EER obtained across the ten trials are summarized.

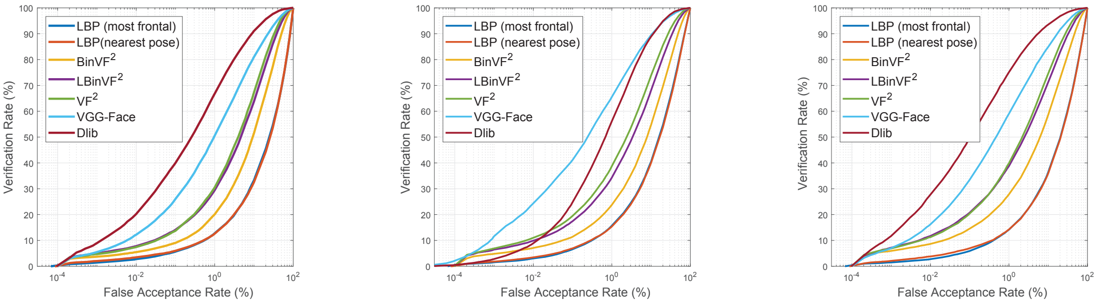
It can be appreciated that, in general, LDA and JB perform better than LMNN. The best results for each evaluated metric learning are obtained by deep learning-based representations, while LBP descriptors achieve a very poor performance, what is consistent with previous results in the standard protocol of the YTF database. We can see from the Figure 1 that for all cases, at lowest FAR ($< 0.01\%$), the TAR of algorithms drops below 20% on average. The lowest EER and top TAR values at different FARs showed in Table 3 are achieved by Dlib+LDA, which are still far from the desired performance.

It should be noticed that REP-YTF verification protocol provides 3,314,989 impostor matches on average, which allows performance evaluation at low FAR values that is not covered under the standard YTF protocol.

5.2. Face Identification

For REP-YTF open/closed-set identification protocols we defined two different scenarios. In the first one (video-to-video comparison), both probe and gallery sets are composed by face videos, while in the second one (video-to-image comparison), the gallery contains a single image per

[2] http://www.cs.tau.ac.il/ wolf/ytfaces/index.html



(a) LMNN (b) JB (c) LDA

Figure 1: Average ROC curves of the different face representation using (a) LMNN, (b) JB and (c) LDA metric learning algorithms for REP-YTF verification protocol. It can be seen that rates drop with decreasing FAR values, and even though the deep learning-based representations perform better, the results at low FARs are poor.

	TAR @ FAR = 0.1%			TAR @ FAR = 1%			EER		
	LMNN	JB	LDA	LMNN	JB	LDA	LMNN	JB	LDA
LBP(most-frontal)	5.98 ± 0.3	6.81 ± 0.2	6.33 ± 0.4	13.19 ± 0.4	16.26 ± 0.5	14.60 ± 0.4	38.01 ± 0.8	32.46 ± 0.4	35.39 ± 0.5
LBP(nearest-pose)	6.47 ± 0.5	7.35 ± 0.4	7.31 ± 0.3	13.10 ± 0.5	15.95 ± 0.6	14.66 ± 0.4	38.32 ± 0.7	32.65 ± 0.6	35.74 ± 0.5
BinVF^2	9.76 ± 0.7	12.35 ± 0.9	15.47 ± 0.9	20.87 ± 0.8	24.62 ± 0.9	28.56 ± 0.7	25.58 ± 0.7	24.73 ± 0.8	23.04 ± 0.5
LBinVF^2	14.88 ± 0.8	18.12 ± 0.7	21.27 ± 0.5	30.41 ± 0.9	35.25 ± 1.0	39.59 ± 0.8	20.14 ± 0.4	18.99 ± 0.9	18.12 ± 0.7
VF^2	14.76 ± 1.0	20.29 ± 0.8	20.84 ± 0.4	32.01 ± 1.5	39.83 ± 1.1	40.68 ± 0.8	19.18 ± 0.7	16.73 ± 0.6	16.37 ± 0.5
VGG-Face	27.33 ± 1.3	**43.04 ± 1.9**	34.38 ± 0.9	51.84 ± 1.3	**66.91 ± 1.4**	59.67 ± 0.6	14.05 ± 1.8	**9.93 ± 0.8**	12.37 ± 1.3
Dlib	**41.50 ± 1.5**	27.64 ± 2.9	**50.70 ± 1.2**	**67.98 ± 1.2**	58.53 ± 2.4	**75.98 ± 0.9**	**9.12 ± 1.5**	10.11 ± 0.1	**7.59 ± 0.4**

Table 3: Performance evaluation for REP-YTF verification protocol, in terms of mean TAR (at FAR values of 0.1% and 1%) and mean EER with their corresponding standard deviations over the 10 trials.

subject and the probe set consists of face videos. For open-set identification we report DIR at rank-1 for 1% and 10% FAR values, respectively, while for closed-set CMC curves at different rank levels for each gallery size are reported.

5.2.1 Video-to-video comparison

Table 4 and 5 show the open-set identification results of the baseline algorithms reported as mean DIR at rank-1 and it corresponding standard deviation across the 10 trials, for 1% and 10% FAR values, respectively. We can clearly observe that the performance of most of the methods is quite low. The best representation for each metric learning at each openness value is highlighted in the tables. Deep learning-based representations achieve the best results. However, the performance is not good enough, which indicates the real challenge of the REP-YTF open-set identification protocol. For example, at FAR = 1%, the top DIRs for the three openness values (0.2, 0.5 and 0.9) are achieved by Dlib+LDA, obtaining 25.97%, 20.12%, and 17.99%, respectively, and the corresponding rates at FAR = 10% are 47.55%, 41.98%, and 39.02%, which also reveal that the open-set identification performance significantly drops for low FAR values. Moreover, we notice that similar to the case of face verification, in general, LDA performs better than LMNN and JB approaches for all the representations. On the other hand, we can see the impact of including impostor comparisons in the identification experiments, which is not usually considered in existing protocols. It can be appreciated that, in general, when the openness value increases, the rates drop, and for the best algorithms, the falls are greater.

Closed-set identification results of the baseline methods for the different gallery sizes are shown in Figures 2, 3 and 4. It can be seen that, also in this scenario, LDA algorithm performs better than JB and LMNN for all the descriptors and deep learning-based representations remain the most discriminative ones. For example, at rank-1, both VGG-Face and Dlib with LDA achieve identification rates higher than 50% for the three gallery sizes. We observed that the recognition performance of all the evaluated methods decreases while increasing gallery set size. For example, for the smallest gallery size (i.e., 200 subjects) the best results at rank-10 range between 80%-90%, while for the largest gallery size (i.e., 533 subjects) range between 70%-85%. All the algorithms, except LBP-based representations, obtained more than 70% identification rates at rank-100 but this is still far from the ideal performance.

5.2.2 Video-to-image comparison

In the case of video-to-image scenario, the LBP-based representations were not included in the comparison since its

	DIR @ FAR = 1%								
	LMNN			JB			LDA		
	Op (0.2)	Op (0.5)	Op (0.9)	Op (0.2)	Op (0.5)	Op (0.9)	Op (0.2)	Op (0.5)	Op (0.9)
LBP(most frontal)	2.31 ± 0.8	1.75 ± 0.3	1.70 ± 0.2	2.79 ± 0.7	2.43 ± 0.5	2.29 ± 0.4	2.24 ± 0.7	2.03 ± 0.4	1.85 ± 0.3
LBP(nearest pose)	2.52 ± 1.0	2.23 ± 0.5	2.34 ± 0.3	2.27 ± 0.8	2.13 ± 0.5	2.09 ± 0.4	2.76 ± 0.7	2.32 ± 0.3	2.29 ± 0.6
BinVF^2	5.44 ± 1.4	4.13 ± 0.7	4.26 ± 0.6	6.84 ± 1.6	5.29 ± 0.5	5.30 ± 0.4	8.36 ± 1.6	6.86 ± 0.7	7.05 ± 0.8
LBinVF^2	7.27 ± 2.1	5.94 ± 0.8	5.82 ± 0.7	8.98 ± 1.9	7.47 ± 1.1	7.29 ± 0.7	10.05 ± 2.1	8.57 ± 0.8	8.18 ± 1.0
VF^2	7.15 ± 1.9	5.75 ± 0.7	5.69 ± 0.8	10.86 ± 2.2	8.47 ± 1.0	8.33 ± 1.0	10.67 ± 2.4	8.47 ± 0.9	8.84 ± 0.9
VGG-Face	12.36 ± 2.9	9.52 ± 1.2	9.18 ± 1.2	**22.83 ± 3.6**	**18.16 ± 1.8**	**16.28 ± 1.5**	14.10 ± 2.4	10.53 ± 1.3	10.25 ± 0.9
Dlib	**19.92 ± 3.6**	**15.03 ± 1.7**	**13.39 ± 1.9**	8.45 ± 3.2	5.67 ± 1.3	4.89 ± 1.5	**25.97 ± 3.0**	**20.12 ± 1.2**	**17.99 ± 1.5**

Table 4: Performance of baseline methods for REP-YTF open-set identification under different openness values, at FAR = 1% in video-to-video scenario. The results are reported as the mean DIR (%) at rank-1 and the corresponding standard deviation over the 10 trials.

	DIR @ FAR = 10%								
	LMNN			JB			LDA		
	Op (0.2)	Op (0.5)	Op (0.9)	Op (0.2)	Op (0.5)	Op (0.9)	Op (0.2)	Op (0.5)	Op (0.9)
LBP(most frontal)	4.84 ± 1.3	3.67 ± 0.5	3.56 ± 0.5	5.32 ± 0.8	4.56 ± 0.7	3.97 ± 0.6	5.29 ± 1.4	4.33 ± 0.7	3.99 ± 0.7
LBP(nearest pose)	5.20 ± 1.0	4.16 ± 0.5	3.89 ± 0.5	5.77 ± 1.0	4.49 ± 0.6	4.19 ± 0.5	6.21 ± 1.4	4.52 ± 0.7	4.40 ± 0.6
BinVF^2	8.60 ± 1.8	7.05 ± 0.4	6.81 ± 0.6	10.95 ± 1.1	8.63 ± 0.6	8.50 ± 0.8	14.26 ± 2.0	11.41 ± 1.1	10.61 ± 1.0
LBinVF^2	13.43 ± 2.2	10.29 ± 1.4	9.90 ± 1.1	16.31 ± 1.1	13.22 ± 1.1	12.16 ± 0.8	19.14 ± 2.1	15.59 ± 1.2	14.97 ± 1.2
VF^2	13.29 ± 2.1	10.24 ± 0.9	10.02 ± 0.9	18.34 ± 2.5	15.11 ± 1.1	14.13 ± 0.7	19.91 ± 3.5	15.58 ± 1.3	14.94 ± 0.8
VGG-Face	29.19 ± 2.2	23.44 ± 0.9	21.77 ± 1.2	**39.38 ± 2.8**	**32.86 ± 1.6**	**30.52 ± 1.9**	32.38 ± 3.2	26.92 ± 1.2	25.12 ± 1.0
Dlib	**37.26 ± 4.1**	**30.88 ± 2.2**	**28.21 ± 2.5**	25.90 ± 3.9	18.07 ± 2.1	16.33 ± 1.9	**47.55 ± 3.1**	**41.98 ± 2.2**	**39.02 ± 1.8**

Table 5: Performance of baseline methods for REP-YTF open-set identification under different openness values, at FAR = 10% in video-to-video scenario. The results are reported as the mean DIR (%) at rank-1 and the corresponding standard deviation over the 10 trials.

	DIR @ FAR = 1%								
	LMNN			JB			LDA		
	Op (0.2)	Op (0.5)	Op (0.9)	Op (0.2)	Op (0.5)	Op (0.9)	Op (0.2)	Op (0.5)	Op (0.9)
BinVF^2	1.77 ± 0.7	1.39 ± 0.4	1.36 ± 0.3	2.80 ± 0.7	2.18 ± 0.4	2.13 ± 0.4	4.49 ± 1.2	3.37 ± 0.6	3.29 ± 0.5
LBinVF^2	3.16 ± 1.1	2.83 ± 0.5	2.69 ± 0.3	5.28 ± 1.4	3.87 ± 0.5	3.58 ± 0.5	6.58 ± 1.5	4.78 ± 0.8	4.53 ± 0.5
VF^2	3.09 ± 0.8	2.61 ± 0.3	2.43 ± 0.4	5.74 ± 1.9	4.81 ± 0.5	4.68 ± 0.6	5.95 ± 1.5	4.92 ± 0.6	4.82 ± 0.7
VGG-Face	6.60 ± 1.9	4.96 ± 1.1	4.81 ± 1.0	**17.33 ± 2.9**	**14.20 ± 2.4**	**13.14 ± 1.1**	10.36 ± 2.0	7.44 ± 0.8	7.01 ± 0.7
Dlib	**9.51 ± 2.6**	**7.93 ± 1.4**	**6.76 ± 1.3**	5.16 ± 2.0	3.68 ± 1.4	3.09 ± 1.1	**16.62 ± 4.2**	**14.26 ± 1.7**	**11.41 ± 1.0**

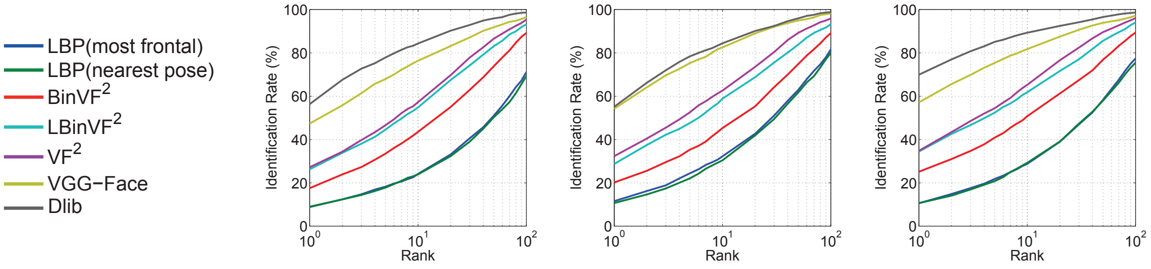
Table 6: Performance of baseline methods for REP-YTF open-set identification under different openness values, at FAR = 1% in video-to-image scenario. The results are reported as the mean DIR (%) at rank-1 and the corresponding standard deviation over the 10 trials.

	DIR @ FAR = 10%								
	LMNN			JB			LDA		
	Op (0.2)	Op (0.5)	Op (0.9)	Op (0.2)	Op (0.5)	Op (0.9)	Op (0.2)	Op (0.5)	Op (0.9)
BinVF^2	4.27 ± 1.1	3.19 ± 0.5	3.09 ± 0.4	5.82 ± 1.1	4.19 ± 0.6	3.99 ± 0.7	8.34 ± 1.1	6.59 ± 1.0	6.08 ± 0.6
LBinVF^2	7.23 ± 1.5	5.64 ± 0.4	5.28 ± 0.5	9.60 ± 2.1	7.40 ± 0.8	7.06 ± 0.6	12.73 ± 2.2	10.03 ± 1.2	9.56 ± 0.7
VF^2	7.81 ± 1.9	6.35 ± 0.8	5.88 ± 0.7	12.33 ± 1.8	9.55 ± 0.9	8.96 ± 1.0	13.58 ± 2.7	10.74 ± 1.3	10.46 ± 0.8
VGG-Face	19.09 ± 2.4	14.32 ± 1.2	13.20 ± 1.0	**32.34 ± 3.0**	**26.93 ± 2.0**	**24.78 ± 1.2**	26.88 ± 2.8	20.65 ± 1.4	19.64 ± 1.0
Dlib	**22.42 ± 3.4**	**18.32 ± 1.8**	**16.24 ± 1.6**	16.91 ± 3.8	12.34 ± 1.8	10.88 ± 2.0	**34.55 ± 4.0**	**30.50 ± 1.3**	**28.01 ± 1.7**

Table 7: Performance of baseline methods for REP-YTF open-set identification under different openness values, at FAR = 10% in video-to-image scenario. The results are reported as the mean DIR (%) at rank-1 and the corresponding standard deviation over the 10 trials.

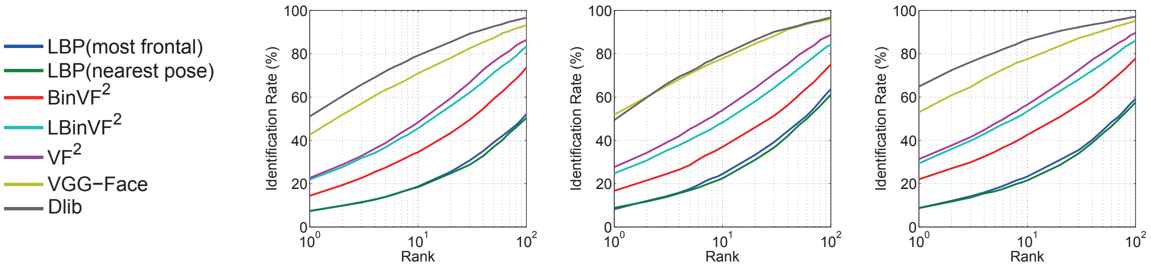
corresponding results presented in Section 5.2.1 are already considering only one frame per video.

The performance obtained by the remaining baseline algorithms at FAR values of 1% and 10% are presented in Table 6 and Table 7, respectively. For each metric learning at each openness value, the best rank-1 DIR is highlighted. It can be seen that when the gallery is composed of still images, the performance is worse than those obtained when both gallery and probe are videos. We suspect that this is because a video contains multiple frames, thus pro-



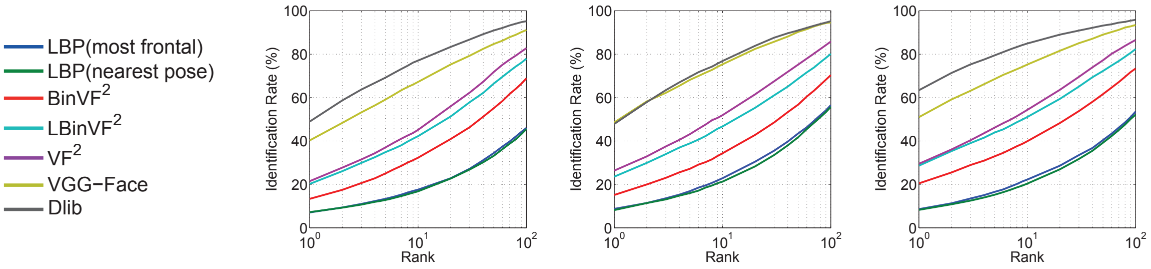
(a) LMNN (b) JB (c) LDA

Figure 2: Average CMC curves of the different face representations using (a) LMNN, (b) JB and (c) LDA algorithms under the REP-YTF closed-set identification protocol for video-to-video scenario. The gallery size is 200 subjects.



(a) LMNN (b) JB (c) LDA

Figure 3: Average CMC curves of the different face representations using (a) LMNN, (b) JB and (c) LDA algorithms under the REP-YTF closed-set identification protocol for video-to-video scenario. The gallery size is 400 subjects.



(a) LMNN (b) JB (c) LDA

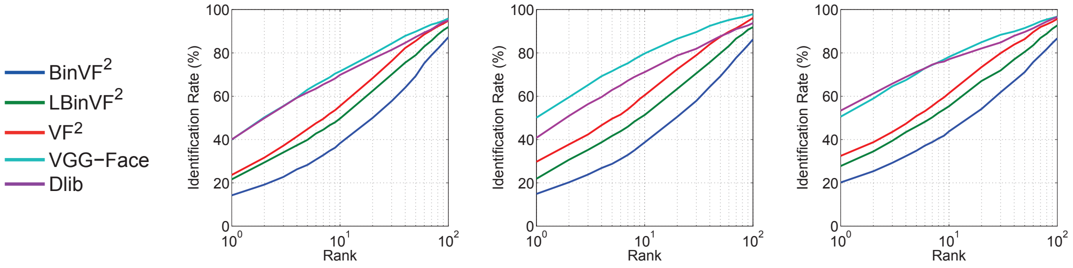
Figure 4: Average CMC curves of the different face representations using (a) LMNN, (b) JB and (c) LDA algorithms under the REP-YTF closed-set identification protocol for video-to-video scenario. The gallery size is 533 subjects.

viding more discriminative information for the recognition. Moreover, deep learning-based representations perform better than the others, however, the best results for each openness value, achieved by Dlib+LDA, are very poor being just 16.62%, 14.26% and 11.41%, respectively, at FAR = 1%, and 34.55%, 30.50% and 28.01%, respectively, at FAR = 10%. Compared with the video-to-video scenario, the best result for each openness value drops more than 10%.

Figures 5, 6 and 7 show the results of the baseline methods under REP-YTF closed-set identification scenario at different gallery sizes. Although the general behavior of the algorithms is similar to that obtained in the video-to-video scenario, the performance drops when the gallery is just a single image per subject. For example, in the case of video-to-video, the best performing algorithm (Dlib+LDA) for the smaller gallery size achieves 70%, 90% and 99% at rank-1, 10 and 100, respectively (see Figure 2c), while in the video-to-image scenario Dlib+LDA obtains 55%, 78% and 97%, respectively, as it can be seen in Figure 5c.

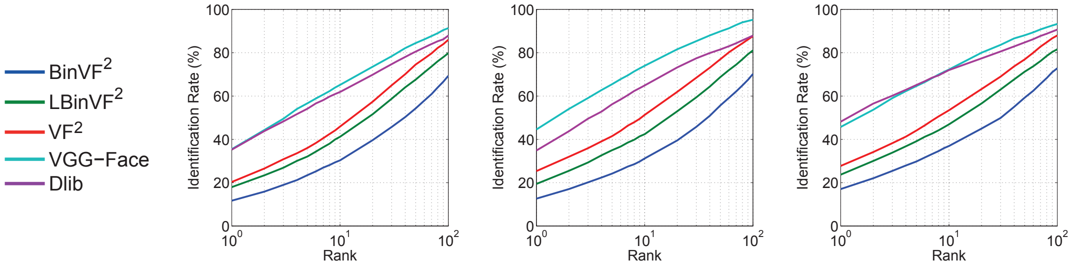
6. Conclusion

In this work, we designed new relevant evaluation protocols for the YouTube Faces database (REP-YTF). The pro-



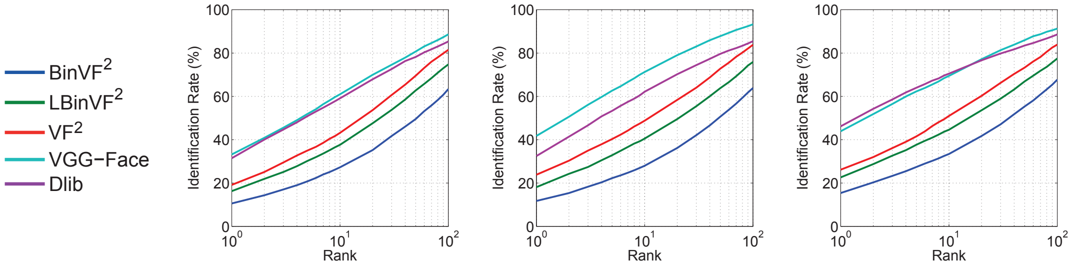
(a) LMNN (b) JB (c) LDA

Figure 5: Average CMC curves of the different face representations using (a) LMNN, (b) JB and (c) LDA algorithms under the REP-YTF closed-set identification protocol for video-to-image scenario. The gallery size is 200 subjects.



(a) LMNN (b) JB (c) LDA

Figure 6: Average CMC curves of the different face representations using (a) LMNN, (b) JB and (c) LDA algorithms algorithms under the REP-YTF closed-set identification protocol for video-to-image scenario. The gallery size is 400 subjects.



(a) LMNN (b) JB (c) LDA

Figure 7: Average CMC curves of the different face representation using (a) LMNN, (b) JB and (c) LDA algorithms algorithms under the REP-YTF closed-set identification protocol for video-to-image scenario. The gallery size is 533 subjects.

posal, which is publicly available, allows the research community to advance face recognition methods under both unconstrained face verification and open/closed-set identification scenarios.

The paper provided an extensive experimental evaluation, including several well-know feature representations with different metric learning algorithms. The results showed that even for the best methods, recognition performances still have a way to go. The benchmark results presented in this paper establish a baseline for evaluating further comparative research on video face recognition.

With this work, we demonstrated that there is room for improvement in the face recognition performance even on well-used benchmarks such as YouTube Faces database. One of the main reasons of that is the lack of appropriate evaluation protocols that model more closely the requirements of operational unconstrained video face recognition scenarios.

7. Acknowledgements

This work was supported in part by FONCICYT Project No. 272727 and the European Union.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. 4

[2] 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 Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8, 2013. 1

[3] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. *Computer Vision–ECCV*, pages 566–579, 2012. 4

[4] C. Ding and D. Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1, 2

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 2, 4

[6] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, and X. Chen. A benchmark and comparative study of video-based face recognition on cox face database. *IEEE Transactions on Image Processing*, 24(12):5967–5981, 2015. 1

[7] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 4

[8] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939, 2015. 1

[9] L. Liu, L. Zhang, H. Liu, and S. Yan. Toward large-population face identification in unconstrained videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(11):1874–1884, 2014. 1

[10] Y. Martínez-Díaz, L. Chang, N. Hernández, H. Méndez-Vázquez, and L. E. Sucar. Efficient video face recognition by using fisher vector encoding of binary features. In *23rd International Conference on Pattern Recognition (ICPR)*, pages 1436–1441, 2016. 4

[11] Y. Martínez-Díaz, N. Hernández, R. Biscay, L. Chang, H. Méndez-Vázquez, and L. Sucar. On fisher vector encoding of binary features for video face recognition. *Journal of Visual Communication and Image Representation*, 51:155–161, 2018. 4

[12] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman. A compact and discriminative face track descriptor. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1693–1700, 2014. 4

[13] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015. 1, 2, 4

[14] J. Phillips. Video challenge problem multiple biometric grand challenge preliminary: results of version 2. In *MBGC 3rd Workshop*, 2009. 1

[15] P. J. Phillips, P. Grother, and R. Micheals. Evaluation methods in face recognition. In *Handbook of Face Recognition*, pages 551–574. 2011. 3

[16] A. T. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1493–1502, 2017. 1, 2

[17] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009. 4

[18] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, et al. Iarpa janus benchmark-b face dataset. In *CVPR Workshop on Biometrics*, 2017. 1

[19] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 1, 2, 4