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

Yoanna Martínez-Díaz, Heydi Méndez-Vázquez, Leyanis López-Ávila
Advanced Technologies Application Center (CENATAV), Havana, Cuba

Leonardo Chang
Tecnológico de Monterrey,
Estado de Mexico, Mexico

L. Enrique Sucar
Instituto Nacional de Astrofísica, Óptica
y Electrónica (INAOE), Mexico

Massimo Tistarelli
University of Sassari,
Sassari, Italy

June 18, 2018

MOTIVATION

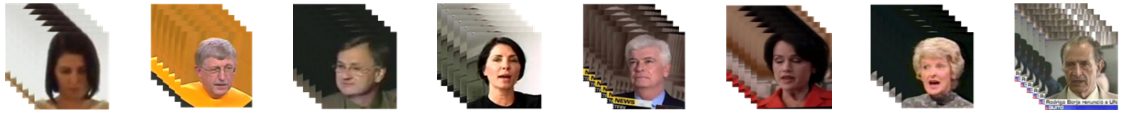
YouTube Face database results

	Method	Accuracy	AUC	EER
2014	VF^2	**84.8**	93	14.9
2018	LBinVF^2	**83.3**	93.2	14.6
2014	DeepFace-single	**91.4**	96.3	8.6
2017	TBE-CNN	**94.9**	-	-
2015	FaceNet	**95.1**	-	-
2016	NAN	**95.7**	98.8	-
2015	VGG-Face	**97.3**	-	2.6
2018	CosFace	**97.6**	-	-
2018	SeqFace	**98.1**	-	-
2016	ResNet-29 (Dlib)	**98.5**	-	-

Is recognition performance saturating for the YouTube Faces database?

Does the standard protocol of the YouTube Faces database capture the requirements of unconstrained scenarios?

YOUTUBE FACES (YTF) DATABASE



- Standard protocol is very limited.
- Only considers the face verification scenario with a reduced number of genuine and impostor comparisons.
- Is not possible to assess the recognition performance at low FAR values.
- Does not support the evaluation of algorithms in the face identification task.
- There are more than 190 videos which are not used.

WHAT TO DO? → **CREATE A NEW DATABASE?**

- Collection and labeling videos of a large number of individuals.
- Design operationally relevant evaluation protocols.

NEW RELEVANT EVALUATION PROTOCOL (REP-YTF)

- It is **clear** and **easy** to understand.
- **A new face verification protocol** that allows the evaluation at **low FAR values**.
- **Open/closed-set identification protocols** considering different gallery sizes, as well as video-to-video and video-to-image comparisons.
- It shows that **face recognition is still an unsolved problem** in the YouTube Faces database.
- It is **publicly available** to encourage and support algorithm development for unconstrained face recognition in videos.

	Standard Protocol	REP-YTF
Use all available data	No	Yes
Closed-set identification protocol	No	Yes
Open-set identification protocol	No	Yes
Face verification protocol	Yes	Yes
# Genuine comparisons	2,500	2,227
# Impostor comparisons	2,500	3,314,989

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

REP-YTF PROTOCOLS

EXPERIMENTAL SETTINGS

- YTF is divided into 10 random trials of training and test sets, ensuring that videos from subjects that are included in the training set are not considered in the test set.
- **Face verification protocol:**
  - On average, **2,277** genuine comparisons and **3,314,989** impostor comparisons not-duplicated are obtained in each trial.
  - It is possible to evaluate face recognition algorithms at low FAR values (e.g., at FAR = 0.1% there are more than 3,300 impostor comparisons available).
- **Face identification:**
  - $\boldsymbol{G}$: gallery set, $\boldsymbol{P_G}$: genuine probe set, $\boldsymbol{P_I}$: impostor probe set
  - This partitioning procedure is repeated three times, varying the openness (Op).
  - Two kinds of gallery are designed (face videos and face image per subject).
- **Closed-set identification protocol**: $\boldsymbol{P_G}$ vs. $\boldsymbol{G}$
- **Open-set identification protocol**: $\boldsymbol{P_G} \cup \boldsymbol{P_I}$ vs. $\boldsymbol{G}$

			# Subjects	# Videos
	Train		395	849
Test	**Verification**		1,200	2,576
	Op (0.2)	G	200	200
		$P \downarrow G$	200	370
		$P \downarrow I$	1,000	2,005
	Op (0.5)	G	400	400
		$P \downarrow G$	400	728
		$P \downarrow I$	800	1,448
	Op (0.9)	G	533	533
		$P \downarrow G$	533	975
		$P \downarrow I$	667	1,068

REP-YTF PROTOCOL

PERFORMANCE METRICS

Open-set Identification

- Detection and Identification rate (DIR)
- False Acceptance Rate (FAR)

Closed-set Identification

- Cumulative Match Characteristic (CMC)

Face Verification

- Receiver Operating Characteristic (ROC) curve
- Equal Error Rate (EER)

BASELINE METHODS

FACE REPRESENTATIONS

- **Local Binary Patterns (LBP) descriptors**
  - LBP most frontal pose
  - LBP nearest pose
- **Fisher vector encoding**
  - VF^2 descriptor
  - $BinVF^2$ descriptor
  - $LBinVF^2$ descriptor
- **Deep convolutional neural networks**
  - VGG-Face
  - ResNet-29 (Dlib)

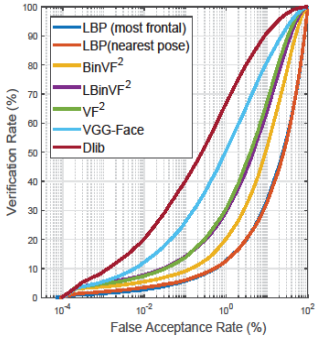
METRIC LEARNING

- **Joint Bayesian (JB)**
- **Large Margin Nearest Neighbor (LMNN)**
- **Linear Discriminant Analysis (LDA)**

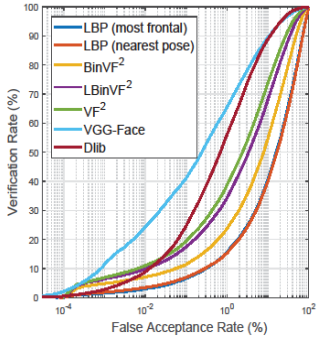
EXPERIMENTAL RESULTS

FACE VERIFICATION

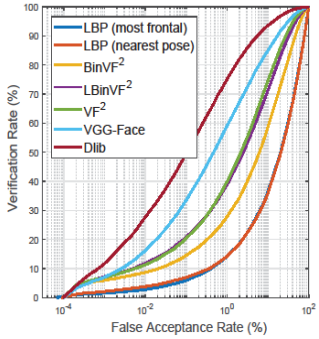
LMNN



JB



LDA



	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
BinVF²	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
LBinVF²	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
VF²	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
ResNet-29 (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**

- In general, LDA and JB perform better than LMNN.
- For each metric learning, deep-based representations achieve the best results.
- The lowest EER and top TAR values at different FAR, are obtained by ResNet-29 (Dlib) + LDA.
- There still much to improve in particular al low FAR!

EXPERIMENTAL RESULTS

(Best results obtained from the experiments)

OPEN-SET IDENTIFICATION

Video-to-video

	DIR @ FAR = 1%			DIR @ FAR = 10%		
	Op (0.2)	Op (0.5)	Op (0.9)	Op (0.2)	Op (0.5)	Op (0.9)
LBP (most frontal) + JB	2.79 ± 0.7	2.43 ± 0.5	2.29 ± 0.4	5.32 ± 0.8	4.56 ± 0.7	3.97 ± 0.6
LBP (nearest pose) + LDA	2.76 ± 0.7	2.32 ± 0.3	2.29 ± 0.6	6.21 ± 1.4	4.52 ± 0.7	4.40 ± 0.6
BinVF² + LDA	8.36 ± 1.6	6.86 ± 0.7	7.05 ± 0.8	14.26 ± 2.0	11.41 ± 1.1	10.61 ± 1.0
LBinVF² + LDA	10.05 ± 2.1	8.57 ± 0.8	8.18 ± 1.0	19.14 ± 2.1	15.59 ± 1.2	14.97 ± 1.2
VF² + LDA	10.67 ± 2.4	8.47 ± 0.9	8.84 ± 0.9	19.91 ± 3.5	15.58 ± 1.3	14.94 ± 0.8
VGG-Face + JB	22.83 ± 3.6	18.16 ± 1.8	16.28 ± 1.5	39.38 ± 2.8	32.86 ± 1.6	30.52 ± 1.9
ResNet-29 (Dlib) + LDA	**25.97 ± 3.0**	**20.12 ± 1.2**	**17.99 ± 1.5**	**47.55 ± 3.1**	**41.98 ± 2.2**	**39.02 ± 1.8**

Video-to-image

	DIR @ FAR = 1%			DIR @ FAR = 10%		
	Op (0.2)	Op (0.5)	Op (0.9)	Op (0.2)	Op (0.5)	Op (0.9)
BinVF² + LDA	4.49 ± 1.2	3.37 ± 0.6	3.29 ± 0.5	8.34 ± 1.1	6.59 ± 1.0	6.08 ± 0.6
LBinVF² + LDA	6.58 ± 1.5	4.78 ± 0.8	4.53 ± 0.5	12.73 ± 2.2	10.03 ± 1.2	9.56 ± 0.7
VF² + LDA	5.95 ± 1.5	4.92 ± 0.6	4.82 ± 0.7	13.58 ± 2.7	10.74 ± 1.3	10.46 ± 0.8
VGG-Face + JB	**17.33 ± 2.9**	14.20 ± 2.4	**13.14 ± 1.1**	32.34 ± 3.0	26.93 ± 2.0	24.78 ± 1.2
ResNet-29 (Dlib) + LDA	16.62 ± 4.2	**14.26 ± 1.7**	11.41 ± 1.0	**34.55 ± 4.0**	**30.50 ± 1.3**	**28.01 ± 1.7**

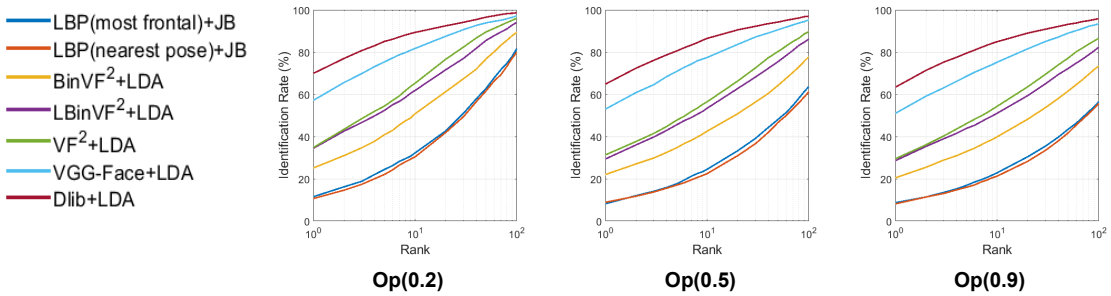
- The best results are obtained by ResNet-29 (Dlib) + LDA, however they are under 50%.
- Deep-based representations are more discriminative.
- LDA performs better than JB and LMNN.
- DIR significantly drops at low FAR values.
- The higher Op value, the lower performance, and for the best methods, the falls are greater.
- Video-to-image scenario seems to be harder than video-to-video scenario.

EXPERIMENTAL RESULTS

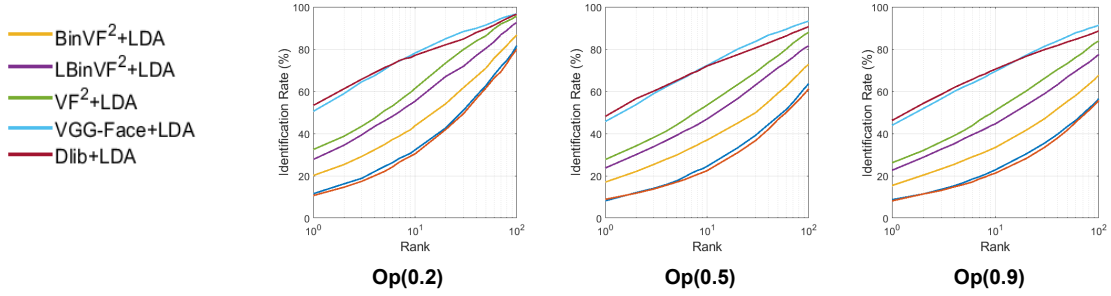
(Best results obtained from the experiments)

CLOSED-SET IDENTIFICATION

Video-to-video



Video-to-image



- Similar behavior to open-set identification but the recognition values are higher.
- The top identification rates at rank-1 range between 40%-75%.
- Near 100% identification rates are obtained at rank-100.

WHY USE REP-YTF?

- Model more closely the requirements of operational **unconstrained scenarios** for video face recognition.
- Allow for evaluation at more operationally relevant points at **low ends of the ROC curve**.
- Support face identification evaluation with **different sizes and types of gallery and openness values**.
- Benchmark results establish a **baseline for evaluating further comparative research on video face recognition** and highlight that recognition performance on the YouTube Faces database still has way to go.
- Show that, by using appropriate evaluation protocols, **there is room for improvement in the face recognition performance** even on well-used benchmarks such as YouTube Faces database.
- A **benchmark toolkit is publicly released** at http://www.cenatav.co.cu/doc/code/REP-YTF.zip

THANKS!

Benchmark toolkit:

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