End-to-End Protocols and Performance Metrics For Unconstrained Face Recognition

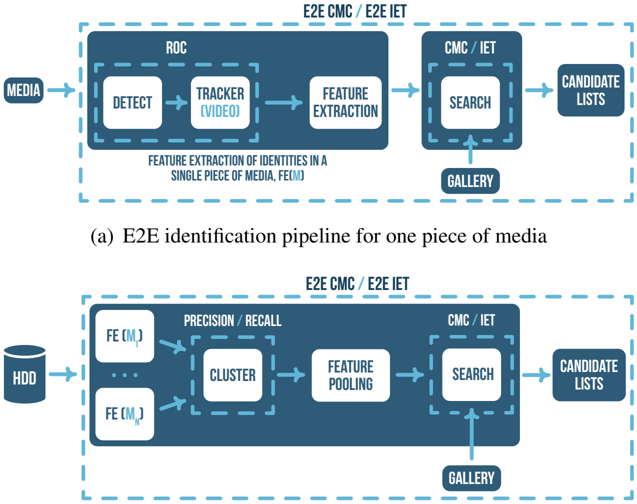
James A. Duncan[1] Nathan D. Kalka[1] Brianna Maze[1] Anil K. Jain[2]
[1]Noblis, Bridgeport, WV, U.S.A.
[2]Michigan State University, East Lansing, MI, U.S.A.
{andrew.duncan,nathan.kalka,brianna.maze}@noblis.org, jain@cse.msu.edu

Abstract

Face recognition algorithms have received substantial attention over the past decade resulting in significant performance improvements. Arguably, improvement can be attributed to the wide spread availability of large face training sets, GPU computing to train state-of-the-art deep learning algorithms, and curation of challenging test sets that continue to push the state-of-the-art. Traditionally, protocol design and algorithm evaluation have primarily focused on measuring performance of specific stages of the biometric pipeline (e.g., face detection, feature extraction, or recognition) and do not capture errors that may propagate from face input to identification output in an end-to-end (E2E) manner. In this paper, we address this problem by expanding upon the novel open-set E2E identification protocols created for the IARPA Janus program. In particular, we describe in detail the joint detection, tracking, clustering, and recognition protocols, introduce novel E2E performance metrics, and provide rigorous evaluation using the IARPA Janus Benchmark C (IJB-C) and S (IJB-S) datasets.

1. Introduction

Traditional face recognition systems operating in verification or identification scenarios perform quite well on constrained media with uniform illumination, pose, and expression. Generally, in these situations there are constraints in place to control the imaging conditions and the behavior of subjects. For instance, at law enforcement booking, airport customs, or border control stations, the operator may provide verbal or visual cues to subjects to limit variations in face capture. These systems perform at or above human levels of performance [12]. However, when capture constraints are relaxed or opportunistic in nature such as in surveillance applications, facial recognition is more challenging.



(a) E2E identification pipeline for one piece of media

(b) E2E identification pipeline for multiple pieces of media

Figure 1. Illustrations of E2E identification pipelines for face recognition. Processing of a single still image or video is shown in (a). Faces are detected or in the case of video, identity tracks are generated, templates are created then searched against a gallery. In (b) multiple still images or videos are processed. The processed media[§]is then clustered to create an identity cluster which is then searched against a gallery. Traditional biometric system performance metrics (ROC, CMC[¶], IET[‖], and Precision/Recall) are indicated with appropriate system component performances. Note that E2E metrics measure the system performance – not individual components.

Several research groups [15, 4, 9, 20, 11, 8] have illustrated that non-ideal factors such as facial expressions, occlusions, non-frontal pose, and low resolution confound face recognition systems. Different strategies for mitigating these challenges have been proposed, but the most

[§] Images or videos.

[¶] Closed set/all probe subjects are in the searched gallery.

[‖] Open set/probe searches are not guaranteed to have a corresponding mate in the searched gallery.

978-1-7281-3640-0/19/$31.00 ©2019 IEEE

Dataset	# subjects	average # images per subject	average # videos per subject	# annotations	avg. fps
IJB-S [8]	202	7	12	>10M	30
MBGC [14]	821	4	5	–	30
UCCS [5]	1,732	approx. 13	N/A	>70,000	1
IJB-A[9]	500	11	4.2	>1.5M	30
IJB-B[20]	1845	6	3.8	>2.3M	30
IJB-C[11]	3,531	6	3	>3.3M	30
LFW[6]	5,749	2	N/A	13,233	N/A
PaSC[2]	265	46 (train + test)	11	248,637	–
Stallkamp et al. [17]	41	–	56	–	–
PubFig [10]	200	294	1.3	N/A	N/A

Table 1. A comparison of unconstrained face still image and video testing datasets. Entries in the table with "–" indicate that the corresponding information was not provided. Note that only the IJB testing benchmarks contain E2E protocols in comparison to all the other benchmark testing datasets.

promising one includes deep convolutional neural networks (DCNN) [13, 16, 18]. DCNNs have become popular due to the widespread availability of large training sets and affordable faster processors (e.g., GPUs). Benchmarks such as the Labeled Faces in the Wild (LFW) [6] and IARPA Janus Benchmarks [9, 20, 11] (example media in Figure 2) have shown that DCNNs significantly outperform prevailing face representation methods and can even approach human-level performance for some benchmarks. These benchmarks, specifically the standard verification benchmark, are no longer challenging.

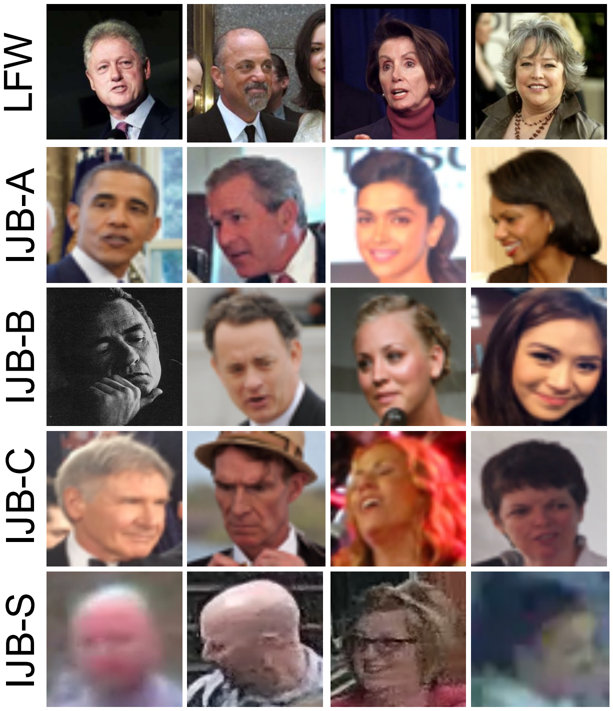


Figure 2. Faces from still images and video frames from LFW, IJB-A, IJB-B, IJB-C, and IJB-S.

However, the problem of 1:N identification, or search, is still challenging for DCNNs. While face verification requires only a single comparison, face search is more demanding as the size of the gallery becomes large (millions of identities) [19]. This problem is made even more challenging in an open-set scenario when probe searches may not have a corresponding mate enrolled in the gallery. The difficulty associated with open-set face recognition protocols is even more challenging when adding face detection evaluation to the pipeline, leading to an E2E recognition protocol. E2E protocol has not received adequate attention in the face recognition literature. Instead, the focus has primarily been on specific modules of the face recognition pipeline such as detection, frontalization, clustering, and/or matching.

A joint detection, including any preprocessing, and open-set search protocol, as illustrated in Figure 1, requires a face recognition system to first detect faces, create a template, and then search against a gallery which may or may not contain a mate for this template. Errors in face detection (e.g., a missed or false detection) will propagate to the recognition stage. This type of protocol resembles the operational use case in law enforcement and intelligence communities where a hard drive is seized containing multimedia (still images and video). The media can then be searched against a watchlist for persons of interest [1].

An even more demanding E2E protocol might take advantage of subject specific modeling where templates are composed of multiple images, video frames, and even sketches of the same identity. This requires detected faces to be clustered first, as illustrated in Figure 1, before searching against a gallery. Errors from clustering will introduce chimeric** templates, or templates composed of faces from different identities which will negatively impact the overall recognition performance.

**Templates created from features of two or more subjects.

In this paper, we address the problem of protocol creation that is even more challenging than the open-set E2E face identification protocols in the IARPA Janus program [7]. In particular, the perceived contributions of this work are three-fold:

1) Defining new E2E performance metrics

2) Benchmarks for VGG and Facenet using the IARPA Janus Benchmark C (IJB-C) and S (IJB-S) datasets

3) An open source reference implementation of the E2E metrics[††]

2. Related Work

Testing datasets generally come with protocols for specific modules such as face detection, representation, verification, or identification. Testing protocols are necessary to promote comparative analysis, benchmark the state-of-the-art, and increase reproducibility and repeatability of empirical results. In the case of face detection, the protocol lists still images or videos that contain faces that must then be detected. In some cases, a subset of the media may not contain any faces. This is the protocol's guidelines for calculating the false detection error rate. Verification protocols provide a list of mated and non-mated template comparisons. Each template is comprised of a face or list of faces as described by the protocol. The latter method is known as subject specific modeling. Similarly, identification protocols list templates for the probe and gallery sets which may also utilize subject specific modeling. The task of identification is more challenging when operating as an open-set problem which is arguably more relevant in applications such as deduplication, searching watchlists, etc. This means that the identities associated with probe media or templates may not have a corresponding identity enrolled in the gallery. This can be achieved by ensuring that a subset of the probe templates do not have a corresponding template in the gallery. The subset of templates is used to calculate the false positive identification rate (FPIR).

Table 1 provides a summary of unconstrained still image and video based testing benchmark datasets available in the public domain. LFW was introduced in 2007 as a "Media in the Wild" dataset that consists mostly of celebrities curated from the web. LFW was released with a single verification only protocol. The number of comparisons in the protocol are small and therefore cannot be used to measure low error rates (i.e. 0.01%) typically of interest to law enforcement and intelligence communities. Furthermore, a commodity face detector was utilized to assist with curation and therefore the difficulty (e.g., related to pose) of the face dataset (e.g. in terms of pose, illumination, expression variations) is relatively low.

LFW was the first and arguably the most influential in the development of unconstrained face recognition algorithms. It also influenced the release of subsequent unconstrained media in the wild datasets such as IJB-A [9], IJB-B [20], PubFig [10], YouTube Faces [21], UCCS [5], and NIST's Point and Shoot (PaSC) [2]. However, as in LFW, the protocols released with these popular datasets do not measure the E2E nature of a biometric recognition system.

IJB-C [11] released in 2018, is a media in the wild dataset which was not biased with a commodity face detector during curation. Instead, identities in the dataset were curated by hand with a human-in-the-loop. The dataset was released with several E2E protocols including (i) joint detection and identification from still images and video frames, (ii) joint detection, tracking, and identification from videos, and (iii) joint detection, clustering, and identification from still images and videos. IJB-S [8] released in 2018, is a test dataset consisting primarily of opportunistic[***] surveillance videos. With the exception of the face detection protocol, all of the protocols released with IJB-S are E2E protocols. Specifically, they are joint detection, tracking, clustering, and open-set identification protocols.

3. Approach

We utilize the E2E protocols released with both IJB-C and IJB-S. The major difference between them is that IJB-S E2E protocols are primarily video-based. We provide a brief summary of E2E testing protocols and present new metrics to support evaluation of the E2E protocols which facilitate finer grained differentiation between competing algorithms performance (e.g., track comparison).

3.1. E2E Protocols

The following is a list and summary of IJB-C and IJB-S E2E protocols that are evaluated throughout the remainder of this paper. An illustration characterizing each protocol can be found in Figure 3. Note that the galleries, G1 and G2, are disjoint to support open-set identification. Each of the two galleries contain roughly half of the identities in each dataset.

1) **E2E still images/frames (IJB-C)**. This is a joint detection and open-set identification protocol. Algorithms are tasked with detecting faces in still images or video frames. Detected faces are searched against curated galleries G1 and G2. This protocol resembles the operational work performed by law enforcement agencies. The protocol includes 136,734 still images and video frames.

[††] This will be released by Fall 2019.

[***] Administrators did not instruct participants to look at cameras or behave in a manner that would be considered unnatural.

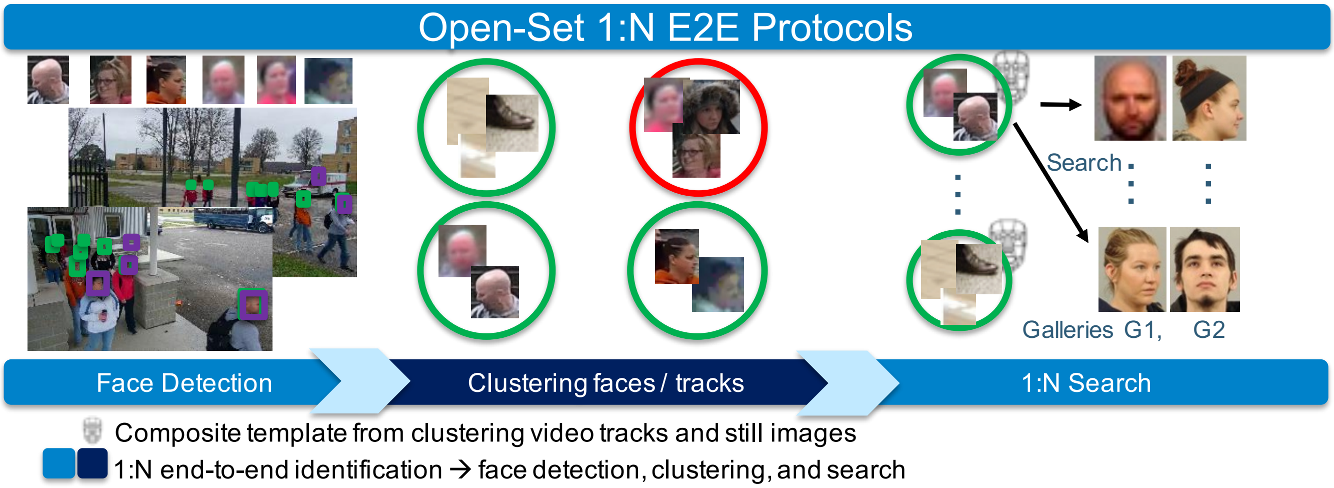

Open-Set 1:N E2E Protocols
Search
Galleries G1, G2
Face Detection
Clustering faces / tracks
1:N Search
Composite template from clustering video tracks and still images
1:N end-to-end identification → face detection, clustering, and search

Figure 3. An illustration depicting the E2E 1:N open set protocols for IJB-C and IJB-S. For joint detection and recognition, faces are detected within a pile of media including both still images and video frames. Each detected face is then searched against two disjoint galleries[§§], G1 and G2. The galleries are disjoint in order to facilitate open-set identification. Joint detection, clustering, and recognition first involve face detection. Each detected face is clustered to create an identity cluster for the purpose of subject specific modeling.

2) **E2E video (IJB-C)**. This protocol supports joint detection, tracking, and open-set identification. The goal is to detect faces, generate identity tracks, and create templates from the identity tracks which are then searched against curated galleries G1 and G2. The protocol includes 11,739 videos.

3) **E2E still image and video (IJB-C)**. This is a joint detection, tracking, clustering, and open-set identification protocol and is the most challenging protocol in IJB-C. The protocol contains 31,415 still images and videos.

4) **Surveillance-to-single/booking (IJB-S)**. Faces are detected, identity tracks are created, and templates are created from identity tracks that are then searched against the curated galleries. The galleries in the single booking protocol consist of a high resolution single mug-shot style photo for each identity. The galleries in the booking protocol consist of multiple high resolution mug-shot style photos for each identity. Both protocols include 398 videos.

5) **UAV Surveillance-to-booking (IJB-S)**. Faces are detected, identity tracks are created, and templates are created from identity tracks that are then searched against the curated galleries. The galleries in this protocol are the same as surveillance-to-booking. The probe video consists of full motion video captured from an unmanned aerial vehicle.

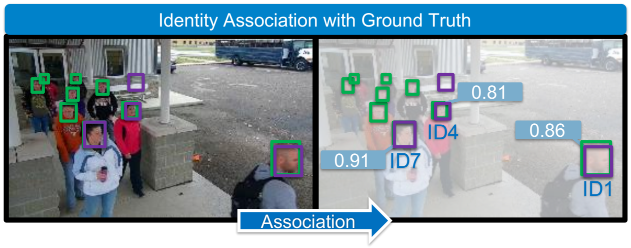


Figure 4. Illustration of identity association through face bounding boxes. The green and purple boxes represent ground truth and predicted boxes, respectively. Associated bounding boxes have an ID number for the subject and corresponding association score.

Protocols 2-5 listed above are expected to utilize subject specific modeling (e.g., tracks from video and identity clusters which may contain still images and tracks). With subject specific modeling, a template is created from multiple faces of the same identity whereas in traditional face recognition tasks, a template is created for each individual detected face.

3.2. Identity Association

All of the protocols described in this section make use of "uncurated" probe media. Labeled data is not provided in the protocols. Instead, the evaluation system performs an association between predicted detections and ground truth detections in the metadata.

Due to the variance of the shape and field-of-view of the predicted bounding boxes generated by face detection algorithms, an area-normalized Jaccard index approach to subject-of-interest association [20] is illustrated in Figure

[§§] The galleries are disjoint to facilitate open-set identification.

4. Predicted bounding boxes are scaled to have the same area as the ground truth bounding box, which is intended to avoid spurious missed detections due to algorithm-specific differences in predicted bounding box size. A predicted detection is considered a true detection whenever (i) ratio of intersection and union of the normalized predicted bounding box and the ground truth bounding box is at least 50% and (ii) the predicted bounding box has not been scaled more than 150%.

3.3. E2E Evaluation Metrics

Given a set of media, $\mathcal{M}$, and set of subjects, $\mathcal{S}$, let all ground truth metadata consist of a media/subject mapping which yields a set of bounding boxes, represented by $\mathcal{N}_{m,s}, m \in \mathcal{M}, s \in \mathcal{S}$. Let subjects of interest be enrolled in galleries $\mathcal{G}_n$ where $\mathcal{G}_n \in \mathcal{S}$. Let the total media-wise occurrences of all subjects (sightings) of interest from gallery $\mathcal{G}_n$ be defined as T. For this set of media, a face recognition algorithm outputs detections $\mathcal{D}_m, m \in \mathcal{M}$ and a candidate list, $c \in \mathcal{C}$ containing scores c_s obtained when comparing a set of detected identities (c_d) in a piece of media (c_m) to a gallery identity ($c_g \in \mathcal{G}_n$) at rank (c_r).

3.3.1 Counting False Positives

The number of false positives is reported given a threshold, t, and gallery, $\mathcal{G}_n$.

$$\mathrm{E2E}_{\mathrm{FP}}(t, \mathcal{G}_n) = \sum_{c \in \mathcal{C}} \sum_{s \in c_d \setminus \mathcal{G}_n} \sum_{c_s >= t} |c_{d=s,r=1}| \qquad (1)$$

3.3.2 False Negative Identification Rate

Given the outputs described by Figure 4, the False Negative Identification Rate ($\mathrm{E2E}_{\mathrm{FNIR}}$) can be evaluated for all searches at all score thresholds, $t \in \mathcal{T}$, for a gallery containing subjects $\mathcal{G}_n$. The minimum value of $\mathrm{E2E}_{\mathrm{FNIR}}$ is determined by the faces missed by a face recognition system's detector, $\mathcal{B}$, as described below:

$$\gamma\left(\mathcal{G}_n\right) = \frac{1}{T} \sum_{m \in \mathcal{M}} \sum_{s \in \mathcal{G}_n} \frac{|\mathcal{B}_{m,s}|}{|\mathcal{N}_{m,s}|} \qquad (2)$$

$\mathrm{E2E}_{\mathrm{FNIR}}$'s variable component can be defined as the sum of ratios of missed detections for an identity to the corresponding number of ground truth detections as described by Equation 3.

$$\eta(t, \mathcal{G}_n) = \frac{1}{T} \sum_{c \in \mathcal{C}} \sum_{s \in \mathcal{G}_n \cap c_d} \sum_{c_s < t} \sum_{m \in c_m} \frac{|c_{d=s,m=m}|}{|\mathcal{N}_{m,s}|} \qquad (3)$$

Finally, $\mathrm{E2E}_{\mathrm{FNIR}}$ is fully defined as follows:

$$\mathrm{E2E}_{\mathrm{FNIR}}\left(t, \mathcal{G}_n\right) = \gamma\left(\mathcal{G}_n\right) + \eta(t, \mathcal{G}_n) \qquad (4)$$

3.3.3 Identification Error Tradeoff Curve

$\mathrm{E2E}_{\mathrm{FNIR}}(t, \mathcal{G}_n)$ plotted against $\mathrm{E2E}_{\mathrm{FP}}(t, \mathcal{G}_n)$ yields the Identification Error Tradeoff Curve.

3.3.4 Cumulative Match Characteristic

The E2E Cumulative Match Characteristic ($\mathrm{E2E}_{\mathrm{CMC}}$) is computed for every mated candidate in each returned candidate list for a given gallery, $\mathcal{G}_n$.

$$\mathrm{E2E}_{\mathrm{CMC}}(r, \mathcal{G}_n) = \frac{1}{T} \sum_{c \in \mathcal{C}} \sum_{s \in \mathcal{G}_n \cap c_d} \sum_{m \in c_m} \sum_{1 < i \leq r} \frac{|c_{d=s,r=i}|}{|\mathcal{N}_{m,s}|} \qquad (5)$$

3.3.5 Subject Cumulative Match Characteristic

The E2E Subject Cumulative Match Characteristic ($\mathrm{E2E}_{\mathrm{SCMC}}$) describes the percentage of unique, mated subjects returned on average from a collection of media for a given rank. Let the probe media be defined by $\mathcal{Z}$ consisting of mated subjects in $\mathcal{G}_n$. Let the subjects at rank $r \in \mathcal{R}$ in media $m \in \mathcal{Z}$ be represented by $\mathcal{E}_{r,m}$. Further, to avoid counting a subject of interest multiple times (due to chimeric templates), consider only the lowest-rank occurrence of a subject. Let $\mathcal{E}_{r,m}$ be governed at all r by: $\mathcal{E}_{r,m} \cap \mathcal{E}_{r+1,m} = \varnothing$ and $\mathcal{E}_{r,m} \cap \mathcal{G}_n \neq \varnothing$. Therefore, $\mathrm{E2E}_{\mathrm{SCMC}}$ is fully defined by as follows:

$$\mathrm{E2E}_{\mathrm{SCMC}}(r, \mathcal{G}_n) = \frac{1}{|\mathcal{Z}|} \sum_{m \in \mathcal{Z}} \sum_{1 \leq i \leq r} |\mathcal{E}_{r,m}| \qquad (6)$$

3.4. Comparison to Traditional 1:N

Traditional 1:N evaluation operates under conventional signal detection theory in that all outcomes must be one of: hit (true positive), miss (false negative), false alarm (false positive), or correct reject (true negative). While the proposed E2E evaluation method maintains these outcomes, it distinguishes itself by the following observations:

- Each outcome is a summation of ratios defined by identities in media. For example, consider two of five possible probe frames of subject s are returned at rank 3:
  - $\mathrm{E2E}_{\mathrm{CMC}}$ will reflect a 40% hit at rank 3.
  - $\mathrm{E2E}_{\mathrm{FNIR}}$ will have a maximum value of 60%, given the selection of any threshold.
- Unlike traditional 1:N evaluation, it is not possible to define a False Positive Identification Rate (FPIR) since an algorithm's face detector can report any number of detections. In the traditional sense, this behavior may allow algorithms to artificially reduce perceived FPIR. Instead, raw false alarm counts are returned for all score thresholds.

- It is possible for a single template to generate hits or misses for multiple subjects. If this happens, the template was generated from the features of multiple identities and is known as a chimeric template. These templates may arise from any combination of impure clusters, subject tracking errors, or bad detections. Consider template t is the average feature vector of subjects s_1 and s_2 and false alarm d_f. Let s_1 be returned at rank 2 and s_2 be returned at rank 10.

  - E2E$_{\text{CMC}}$ will reflect appropriately-weighted hits at rank 2 for subject s_1 and at rank 10 for subject s_2.

  - E2E$_{\text{FP}}$ will reflect 1 false alarm at an appropriately-selected score threshold due to d_f.

  - E2E$_{\text{FNIR}}$ will reflect appropriately-weighted misses (between 0 and 2) depending on the selected score threshold.

4. Experimental Results

Baseline results are presented for the datasets and protocols described in section 3 with the exception of the IJB-C E2E video protocol since this is a subset of the E2E still image and video protocol. For the tested protocols, the multi-task cascaded convolutional neural network (MTCCN) [22] algorithm is utilized as a face detector. Detected faces are then encoded and clustered using DBSCAN [3]. For encoding and feature extraction we report performance from an implementation of Google‘s FaceNet[†††], which was shown to achieve a 98.7% accuracy on LFW. Finally, we also report performance of a VGG CNN [13] model. To handle multi-image templates, a single feature vector was composed using a weighted average of the images in the template such that all frames belonging to the same identity within a video have a combined weight equal to a single still image.

4.1. IJB-C

The IJB-C E2E protocol results using VGG and FaceNet are plotted along with curated results for comparison. Curated results do not require association, and make use of ground truth bounding box information and clustering for subject specific modeling. The curated results serve as a loose upper bound on performance for the E2E protocols for the specific combination of detector, clustering, and recognition algorithms.

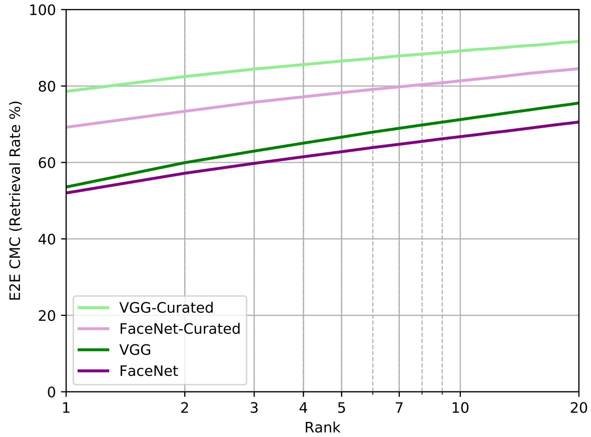


Figure 5. Average E2E closed-set performance across gallery sets G1 and G2 for the still images and frame identification protocol.

4.2. E2E Still Images and Frames

Figure 5 provides the E2E closed-set performance for both VGG and FaceNet algorithms on the still images and frames protocol. As expected, curated VGG performance is roughly 30% higher at rank 1 in comparison to E2E VGG performance. Similarly, curated FaceNet performance is roughly 20% higher than its non-curated counterpart. This not only highlights the difficulty associated with the E2E protocol but also illustrates that subject specific modeling can provide a significant performance gain with accurate clustering.

Figure 6 characterizes E2E open-set identification performance on the still images and frames protocol. Curated VGG and FaceNet provide a lower FNIR at a lower number of false positives in comparison to their non-curated counterparts indicating that they provide significantly better open-set performance for this protocol. Again, the curated results are provided as they serve as a rough upper bound on E2E performance.

4.3. E2E Still Images and Videos

Figure 7 characterizes E2E closed-set performance for both VGG and FaceNet algorithms on the still images and videos protocol. Both baseline algorithms perform substantially below their curated counterparts. VGG performance is roughly 5% below FaceNet. We believe this difference can be attributed to detections that were discarded by our implemention of the MTCNN variant associated with VGG.

4.4. IJB-S

The IJB-S E2E protocol results for FaceNet and VGG are summarized in Table 2. Detection and rank 1 candidate samples are shown in Figure 9. Overall, the UAV protocol

[†††] The implementation can be found at https://github.com/davidsandberg/facenet

	Surv-to-Single		Surv-to-Booking		UAV-to-Booking			Surv-to-Single	Surv-to-Booking	UAV-to-Booking
	Rank 1	Rank 10	Rank 1	Rank 10	Rank 1	Rank 10		FP=10K	FP=10K	FP=10K
VGG	13.82% / 38.19%	25.45% / 58.79%	15.19% / 29.25%	26.57% / 58.47%	0.24% / 1.52%	0.39% / 8.70%		98.92%	93.98%	N/A
FaceNet	5.5% / 15.80%	13.74% / 31.22%	5.98% / 16.85%	14.29% / 32.47%	0.09% / 3.33%	0.35% / 5.18%		97.76%	96.65%	N/A

Table 2. IJB-S closed-set and open-set performance for VGG and FaceNet algorithms for Surveillance-to-Single, Surveillance-to-Booking, and UAV-to-Booking protocols. The E2E CMC/SCMC retrieval rate is reported for rank 1 and rank 10. The FNIR is reported when the # of false positives is 10K. Open-set metrics are not reported for either algorithm on the UAV protocol due to the limited number of probe samples.

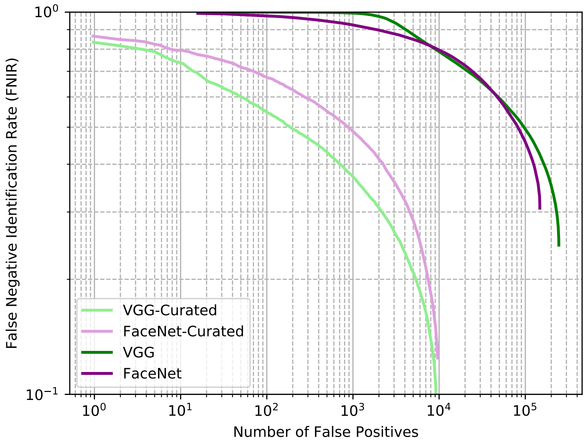


Figure 6. E2E open-set performance across gallery sets G1 and G2 for the still images and frames identification protocol.

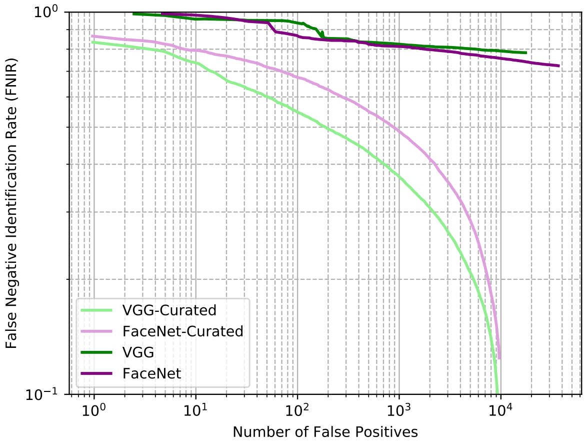


Figure 8. E2E open-set performance across gallery sets G1 and G2 for the still images and frames identification protocol.

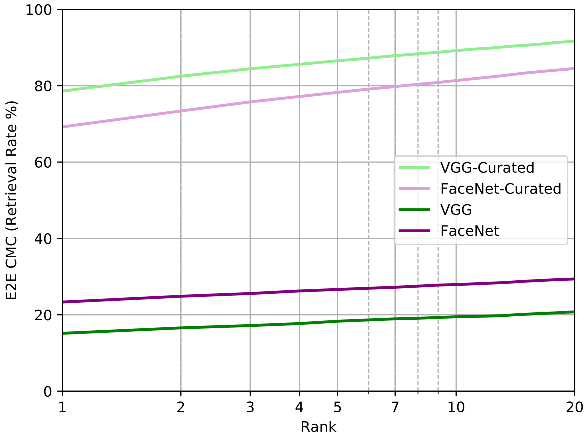
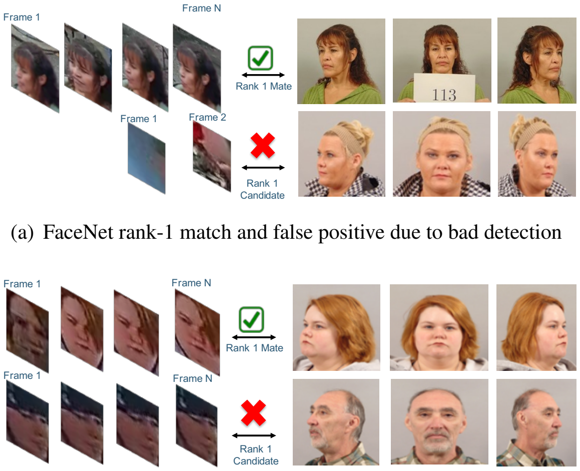


Figure 7. Average E2E closed-set performance across gallery sets G1 and G2 for the still images and video identification protocol.



(a) FaceNet rank-1 match and false positive due to bad detection

(b) VGG rank-1 match and rank-1 false positive

Figure 9. Detection and rank 1 candidate sample faces for FaceNet and VGG on IJB-S.

is clearly more challenging than the other surveillance protocols as the E2E retrieval rates for both VGG and FaceNet are significantly lower in comparison to their respective performance on the other protocols.

5. Summary and Conclusion

We have presented a challenging protocol for face recognition by expanding upon the open-set E2E face identification protocols created for the IARPA Janus program. E2E protocols are more challenging than their traditional counterparts as errors made at early stages in the recognition pipeline will propagate to subsequent stages. In particular, we introduce new metrics for evaluating joint face detection, tracking, clustering, and recognition protocols utilizing the E2E protocols released with both IJB-C and IJB-S benchmarks. Notably for closed set metrics, we introduce the E2E cumulative match characteristic retrieval rate and the subject cumulative match characteristic retrieval rate. We introduce the E2E false negative identification rate and number of false positives expressed through the IET for open-set metrics. We illustrate the new metrics by providing new baseline results with state-of-the-art DCNN implementations of FaceNet and VGG utilizing an MTCNN face detection algorithm.

Our experimental results characterize the challenging nature associated with E2E protocols. In particular, both VGG and FaceNet algorithms perform well under their approximate upper bounds on performance demonstrated by their curated counterparts. Our results demonstrate that the VGG algorithm outperforms FaceNet across the board in both closed-set and open-set protocols with the exception of the IJB-C E2E still images and videos protocol. This result is attributed by low confidence face detections that were discarded prior to clustering.

Acknowledgements

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI) and the Intelligence Advanced Research Projects Activity (IARPA), via FBI Contract #DJF-16-1200-G-0009392. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] 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. *IEEE Transactions on Information Forensics and Security*, 9(12):2144–2157, 2014. 2

[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 BTAS*, pages 1–8, 2013. 2, 3

[3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996. 6

[4] P. J. Grother, M. L. Ngan, and K. K. Hanaoka. Ongoing face recognition vendor test (frvt) part 2: Identification. Technical report, 2018. 1

[5] M. Günther, P. Hu, C. Herrmann, C. H. Chan, M. Jiang, S. Yang, A. R. Dhamija, D. Ramanan, J. Beyerer, J. Kittler, et al. Unconstrained face detection and open-set face recognition challenge. *arXiv preprint arXiv:1708.02337*, 2017. 2, 3

[6] 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. 2

[7] IARPA. Janus. `https://www.iarpa.gov/index.php/research-programs/janus`. 3

[8] N. D. Kalka, B. Maze, J. A. Duncan, K. O'Connor, S. Elliott, K. Hebert, J. Bryan, and A. K. Jain. IJB–S: IARPA Janus surveillance video benchmark. In *IEEE BTAS*, 2018. 1, 2, 3

[9] 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 unconstrainepd face detection and recognition: IARPA Janus Benchmark A. In *IEEE CVPR*, pages 1931–1939, 2015. 1, 2, 3

[10] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *IEEE ICCV*, Oct 2009. 2, 3

[11] B. Maze, J. Adams, J. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. IARPA Janus Benchmark–C: Face dataset and protocol. In *ICB*, 2018. 1, 2, 3

[12] 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. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(9):1642–1646, Sept. 2007. 1

[13] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 2, 6

[14] P. J. Phillips et al. Overview of the multiple biometrics grand challenge. In *IEEE ICB*, pages 705–714, 2009. 2

[15] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998. 1

[16] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015. 2

[17] J. Stallkamp, H. K. Ekenel, and R. Stiefelhagen. Video-based face recognition on real-world data. *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007. 2

[18] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE CVPR*, pages 1701–1708, 2014. 2

[19] D. Wang, C. Otto, and A. K. Jain. Face search at scale. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1122–1136, 2017. 2

[20] 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 CVPR Workshop on Biometrics*, July 2017. 1, 2, 3, 4

[21] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE CVPR*, pages 529–534, 2011. 3

[22] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016. 6