To appear in the Proc. IEEE BTAS, 2013

Video-to-Video Face Matching: Establishing a Baseline for Unconstrained Face Recognition

Lacey Best-Rowden[1], Brendan Klare[2], Joshua Klontz[2], Anil K. Jain[1]
[1]Michigan State University, East Lansing, MI, U.S.A.
[2]Noblis, Falls Church, VA, U.S.A.
bestrow1@cse.msu.edu;{brendan.klare,joshua.klontz}@noblis.org;jain@cse.msu.edu

Abstract

Face recognition in video is becoming increasingly important due to the abundance of video data captured by surveillance cameras, mobile devices, Internet uploads, and other sources. Given the aggregate of facial information contained in a video (i.e., a sequence of face images or frames), video-based face recognition solutions can potentially alleviate classic challenges caused by variations in pose, illumination, and expression. However, with this increased focus on the development of algorithms specifically crafted for video-based face recognition, it is important to establish a baseline for the accuracy using state-of-the-art still image matchers. Note that most commercial-off-the-shelf (COTS) offerings are still limited to single frame matching. In order to measure the accuracy of COTS face recognition systems on video data, we first investigate the effectiveness of multi-frame score-level fusion and analyze the consistency across three COTS face matchers. We demonstrate that all three COTS matchers individually are superior to previously published face recognition results on the unconstrained YouTube Faces database. Further, fusion of scores from the three COTS matchers achieves a 20% improvement in accuracy over previously published results. We encourage the use of these results as a competitive baseline for video-to-video face matching on the YouTube Faces database.

1. Introduction

The increasing ubiquity of surveillance imaging devices offers a promising avenue to combat acts of terrorism and crime. Tragic events such as the Boston bombings[1] in 2013 and the 2011 riots in London[2] have drawn attention to the potential role that surveillance cameras and video-based face recognition can play in identifying perpetrators of such criminal acts. Unfortunately, identification technology has not quite lived up to expectations in such instances. A lack of robustness to classic challenges in pose, illumination, and expression are the prevailing explanations as to why face recognition has limited performance in these unconstrained scenarios. However, equally important is the inability of state-of-the-art technology to leverage the additional temporal information available in sequences of face images (*i.e.*, videos). In order for researchers to improve upon these limitations, we first need to determine the accuracy of deployable technology (*i.e.*, a COTS face matcher) in matching unconstrained faces in video data. Without such a baseline, we cannot determine if meaningful progress is being made in video-based face recognition.

To date, the majority of face recognition research has focused on improving the ability to match static (*i.e.*, still) face images. This is purportedly due to several factors, including (i) the need to constrain the face recognition problem, (ii) computational constraints, and (iii) the large amount of legacy still face images (*e.g.* id cards, mug shots). However, today many of these factors are no longer limiting: still image face recognition has witnessed an exponential decrease in error rates [8], distributed computing readily supports processing a very large number of images, and low cost digital cameras are continuously acquiring an abundance of video data.

Face recognition in video has many applications when applied in a security setting (see Fig. 1). Often, the first step is to perform re-identification, where a collection of videos is cross-matched to locate all occurrences of the person of interest (Fig. 1c). For example, in the recent Boston bombing incident, all available videos (both CCTV and amateur video) needed to be cross-matched to find all instances of the suspected attackers. In turn, the collection of videos of a subject can be used to query legacy face image databases

[1]http://articles.washingtonpost.com/2013-04-16/world/38587918_1_charge-richard-des-lauriers-boston-marathon-explosive-devices
[2]http://latimesblogs.latimes.com/technology/2011/08/london-riots-facial-recognition-tech-being-used-by-police.html

(a) Still-to-Still (b) Video-to-Still (c) Video-to-Video

Figure 1. Face recognition can generally be categorized into one of the following three scenarios based on the characteristics of the image(s) to be matched. (a) Still-to-still image matching is perhaps the most common scenario and is used in both constrained and unconstrained applications. (b) Video-to-still image matching occurs when a sequence of video frames is matched against a database of still images (*e.g.*, mug shots or driver license photos). (c) Video-to-video matching, or re-identification, is performed to find all occurrences of a subject within a collection of video data. Re-identification is generally a necessary pre-processing step before video-to-still image matching can be performed. In this work, we generate a baseline accuracy using commercial face matchers for video-to-video face matching.

in an attempt to identify the subject (Fig. 1b).

In this paper, we study the problem of video-to-video face matching in order to gain a better understanding of state-of-the-art recognition accuracies of COTS face matchers. Through studying the video-based face recognition problem, the methods proposed in this paper can be readily applied in any operational setting using existing COTS matchers (as opposed to dedicated video-based algorithms). Thus, the results provided in this paper offer researchers and practitioners a better understanding of how accurately video face data can be recognized using off-the-shelf technology. For example, we will demonstrate that on an unconstrained, public-domain, video-based face recognition dataset, the highest accuracies previously reported in the research literature are 20% lower than those achieved by our use of static COTS matchers. In turn, subsequent research on video-based face recognition algorithms should demonstrate the ability to improve upon such baseline accuracies.

The contributions of this paper can be summarized as follows. (i) A framework for applying COTS face recognition algorithms to video-based data is provided. (ii) The impact and consistency of different match score fusion rules are studied for several commercial matchers, providing guidance on how to best consolidate frame by frame face match scores across the video. (iii) The performance of COTS algorithms is studied with respect to quality-based key-frame subset selection. (iv) An order of magnitude decrease in error rates is achieved on the YouTube Faces Database [24], respective to the best accuracy currently reported in the literature.

2. Understanding Video-based FR

In this section we provide a brief overview of video-based face recognition methods with an intent to highlight the merits of different approaches and motivate the need for a state-of-the-art baseline. For a more in depth list of video-based face recognition algorithms, readers are referred to the recent survey on this topic [2].

2.1. Prior Work

Video-based face recognition approaches have been organized into the following two categories [2] based on how they leverage the multitude of information available in a video sequence: (i) sequence-based, and (ii) set-based. At a high-level, what most distinguishes these two approaches is whether or not they utilize temporal information.

Sequence-based approaches consider all detected faces based on their temporal ordering. For example, Zhou *et al*. combined both face tracking and face recognition into a single framework, which allowed the inter-frame dynamics to be exploited during the recognition process [26]. However, sequence-based methods are specialized matchers that cannot be readily deployed in operational scenarios. See [2] for more details about sequence-based methods.

Set-based approaches to video-based face recognition consider all the available frames of a subject's face as an unordered set. Such methods have been further organized into approaches that fuse the available information prior to matching and those that fuse information after matching [2]. Methods that fuse information prior to matching will generally output either a feature vector representation or a single face image. For example, manifold-based methods project the set of face images onto a manifold within a feature space, which in turn facilitates matching within the feature space [14, 22]. Manifold methods are similar to sequence-based methods in that they require specialized matching algorithms. Both super resolution methods [1] and 3D modeling-based methods [18] output a single face image that in turn can be matched with a face recognition system. Thus, while such synthesis-based methods attempt to solve a difficult generative modeling task, these methods are compatible with COTS face recognition engines. A few commercial solutions are available for such synthesis methods, though they are only semi-automated and hence more relevant to forensic applications.

Finally, set-based methods that fuse information after the face matching process seek to combine the match scores from static face matchers into a single score. While

Table 1. Characteristics of popular face video datasets and reported identification accuracies.

Database	Settings	No. of Subjects	No. of Videos	Accuracy
Motion of Body (MoBo) [7]	Treadmill walking: slowly, quickly, on incline, or with a ball	25	150	98.8% [16]
Face in Action (FIA) [6]	Variations in expressions and orientations; indoor/outdoor	221	n/a	99% [18][a]
1st Honda/UCSD [13]	Staged head rotations and expressions	20	75	99% [20]
MBGC [19]	Walking, activity, conversation; standard and high resolutions	821	3,764	see [19]
YouTube Celebrity [11]	Unconstrained, many same-subject tracks from the same video	47	1,910	78.9% [25]
YouTube Faces [24]	Unconstrained	1,595	3,425	23.34%, 38.4%[b,c]

[a] Authors used an indoor subset of FIA [b] TAR @ FAR = 0.1%, 1.0% [15] [c] We demonstrate that COTS face matchers achieve higher accuracy

these methods cannot explicitly leverage the temporal information available (unlike sequence-based methods), nor can they leverage redundant (*i.e.*, confirmatory) information (unlike feature-based fusion methods), these set-based methods have the notable advantage of being able to easily leverage proven COTS face matchers.

2.2. Representative Baselines

The afore-discussed approaches to video-based face recognition can also be categorized based on their technology readiness level. That is, all video-based face recognition algorithms are not equal in their capabilities. Accuracy aside, methods that can quickly be integrated into operational environments are preferable to those that have only been demonstrated in proof of concept (*i.e.*, academic) implementations. Because specialized matching algorithms are not readily usable in operational environments, the first question that must be answered for these methods is, how much, if any, does such a specialized matcher improve upon deployable algorithms? This question is often not properly addressed, making it unclear if such proposed algorithms are worth the effort to engineer into operational systems.

Along these lines, while many previously proposed video-based face matching algorithms have been compared to frame-based static matchers, the static (or still image) matchers chosen were often not representative of state-of-the-art matchers. For example, only a few previous publications have demonstrated improved recognition accuracy using COTS face matchers as a baseline (*e.g.*, [17, 18]).

Thus, the motivation for this work is as follows: we provide a baseline accuracy for video-based face recognition by using state-of-the-art COTS face matchers. This way, the merits of specialized face recognition systems for video data subsequently proposed in the literature can be properly evaluated based on their ability to improve over this baseline accuracy.

2.3. Video Face Databases

A summary of the common public domain databases used to evaluate video-based face recognition algorithms can be found in Table 1. Of particular interest for these datasets is the number of subjects available and whether or not the activities of the subjects were constrained or unconstrained. Notably, the YouTube Faces (YTF) database [24] contains the largest number of subjects and has the highest degree of unconstrained behavior. While the MBGC video data [19] also has strong relevance to the problem being studied, we decided to use the YTF database for the following reasons: (i) it contains the largest number of subjects, (ii) the actions of the subjects are naturally varied (as opposed to performing prescribed actions), (iii) the YTF database is easier to acquire (thus allowing the baselines to be used by the research community at large), and (iv) all subjects in the YTF database also have still images available in the Labeled Faces in the Wild (LFW) database [9] (thus allowing baselines to be compared to the video to still image matching scenario).

3. Benchmarking Video-to-Video Matching

With the exception of Cognitec's *FaceVACS-VideoScan*[1], most commercial face recognition systems are only capable of still image enrollment and matching. Therefore, practical application of COTS matchers for video-based face recognition generally involves formulating the problem in terms of single frame matching.

The first step in applying static face matchers to video data is to detect the faces that are present in the video. Here, any face detection algorithm can be applied, including the underlying matcher's face detector. The output from this step will be a set of face locations and corresponding time stamps. From this set of face locations and times, different *face tracks* are formed, where a face track is any sequence of extracted faces which can be assumed to be of the same person (*i.e.*, consecutive frames of the same face).

This work is not concerned with challenges in face track formation. Instead, our focus is on how to best match face tracks using black box COTS face matchers. However, the topic of face detection and tracking in video is still an active research area (see [2] for additional information).

[1] http://www.cognitec-systems.de/FaceVACS-VideoScan.20.0.html

Figure 2. Example images from face tracks of two subjects. Top two and bottom two rows are face tracks from the same subject.

A face track can be represented as a set of images, $U = \{u_1, u_2, ..., u_a\}$, where u_i is the i-th face detected/extracted from a consecutive frames of a video. Given two face tracks U and V, each containing a and b faces, respectively, we can apply any static-image face matcher to all frame-to-frame pairs (u_i, v_j) to obtain $a \times b$ similarity scores $s(u_i, v_j)$. Hence, the comparison of two face tracks results in a similarity matrix $S(U, V)$. Under the assumption that each face track contains faces of a single person, we wish to recognize if two face tracks represent the same identity. Whether for face identification or face verification, *the track similarity matrix needs to be resolved to a single similarity score* indicating the overall similarity between the two identities present in the two face tracks.

4. Multi-Frame Fusion

As discussed in the previous section, a face track is a set of multiple face samples of the same subject. As with biometric fusion from multiple sensors (*e.g.*, near-infrared and visible face images) or modalities (*e.g.* face and iris), fusion from multiple frames of a video track can be conducted at the feature, score, decision, or rank levels [10]. 3D modeling (*e.g.*, [17, 4]) and super-resolution (*e.g.*, [23]) are examples of feature-level fusion. Comparatively, fusion at the rank, decision, and score levels is simpler, as it can be applied to the outputs of existing matchers. Rank-level fusion is a popular approach for video-based face recognition. For example, with the majority voting rule, the identity present in a face track is determined by the majority vote amongst all frame-to-frame identity decisions. Rank-level fusion does have notable drawbacks: it is only applicable in identification paradigms and has an additional computational cost of sorting to determine ranks.

In this work, we focus on score-level fusion, which supports both verification and identification matching paradigms. Furthermore, score-level fusion provides more information than rank and decision levels. Kittler *et al*. used Bayesian estimation theory to formulate the verification decision from multiple samples [12]. Based on the assumption that the posterior probability estimates from single samples are corrupted by noise, they show that a combination of these estimates by averaging, max, min, and median rules reduces error theoretically and improves performance empirically over using a single sample. In the video-to-video matching framework, the max and mean rules are given by Eqs. 1 and 2, respectively.

$$s(U, V) = \max_{i,j} s(u_i, v_j) \tag{1}$$

$$s(U, V) = \frac{1}{ab} \sum_{i=1}^{a} \sum_{j=1}^{b} s(u_i, v_j) \tag{2}$$

The min and median rules are formulated similar to the max rule. Note that the mean rule is also referred to as the sum rule in the literature. We use the term "mean" to emphasize that the sum has to be normalized by the number of frames in tracks U and V. This is because in unconstrained video-to-video matching, the video sequence lengths can vary. By contrast, in traditional sum fusion, the number of sources of evidence (*e.g.*, sensors, modalities, matchers) is a constant, so this normalization is not necessary.

4.1. Quality-based Fusion

The performance of static COTS matchers typically depends on the quality of the input face images. Quality can be a measure of *faceness* (the confidence or reliability that the detected object is a face [20]), image sharpness, illumination conditions, facial pose, etc., or a combination of multiple attributes. Within a given face track, face images often span a variety of such quality measures. In order to boost the accuracy of matching a face track, good quality frames should have more impact on the final matching result than low quality frames. Selecting a subset of the highest quality frames from each face track prior to matching has been addressed in, for example, [20] and [18]. In this paper, we evaluate the performance of COTS matchers when key frames are selected based on the highest face confidence and most near-frontal pose.

4.2. Fusion of Multiple Matchers

In addition to combining the evidence across multi-frame face tracks to determine the verification decision, we also consider fusion of multiple matchers. There are two pipelines for how this can be done: 1) fusion of multiple matchers can be applied to the frame-to-frame scores (*i.e.*, *before* multi-frame score fusion across two videos), or 2) fusion of multiple matchers can be applied to the video-to-video scores (*i.e.*, *after* multi-frame score fusion across two videos). We denote these two approaches as Multi-Matcher Multi-Frame (MMMF) and Multi-Frame Multi-Matcher (MFMM) fusion. To fuse the similarity scores by multiple matchers, the max, min, median, and sum rules

are applied in conjunction with these multi-matcher fusion schemes.

For example, suppose the mean rule is used for multi-frame fusion of $a \times b$ scores (recall a and b are the number of frames in the two tracks U and V, respectively), and the max rule is used for multi-matcher fusion of m scores obtained from m matchers. The final video-to-video similarity scores for the MMMF and MFMM fusion pipelines are given by Eqs. 3 and 4, respectively.

$$s_{mf}(U,V) = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\max_{k=1,2,\ldots,m} s_k(u_i,v_j)\right) \quad (3)$$

$$s_{fm}(U,V) = \max_{k=1,2,\ldots,m}\left(\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} s(u_i,v_j)\right) \quad (4)$$

Because the m frame-to-frame scores, $s_k(u_i, v_j)$, or m video-to-video scores, $S_k(U, V)$, obtained from different COTS matchers may fall in different ranges and/or follow different statistical distributions, score normalization is necessary. We utilize four simple transformation-based normalization schemes: min-max, z-score, median, and tanh (see [10]).

5. Experiments

Data: We evaluate the performance of three different COTS matchers on a subset of the YouTube Faces (YTF) database. The 1,447 subjects and 3,226 videos in this subset are those included in the evaluation protocol provided by Wolf *et al*. [24]. The number of videos per subject ranges from one to six, with an average of 2.15 videos. The longest and shortest videos contain 48 and 2,157 frames, respectively, with an average of 182 frames per video.

All videos were downloaded from YouTube by searching the names of the same subjects in the Labeled Faces in the Wild (LFW) database [9]. Faces were detected with the Viola-Jones face detector [21]. Videos with faces "stably" detected in at least 48 consecutive frames were included. A face detection was considered stable if the Euclidean distance between its center and the center of the preceding detection was less than 10 pixels. In addition to providing the raw frames of each video at 24fps, Wolf *et al*. provide each subject's video in the same manner as face tracks. Images were cropped to 300×300 pixels and aligned after expanding the detected bounding boxes by a factor of 2.2 [24]. The task at hand is matching face tracks to face tracks, thus replicating the cross video matching scenario that is of strong interest for re-identification in surveillance videos.

We directly feed the 300×300 images to the COTS matchers, whereas [24] and [15] further crop the aligned images to the central 100×100 pixels of the face. We allow the COTS matchers to perform their own enrollment (*i.e*., face and landmark detection) on the raw face track images because: i) COTS matchers may fail to enroll a face if it has been tightly cropped, ii) the center 100×100 pixels may not be the best cropping for a given matcher, and iii) a rough initial cropping after detection is more representative of operational settings. Note that there is no scale difference (*i.e*., the interpupillary distances remain the same) between the images used here and those used by [3, 15, 24]. However, [5] resizes the cropped faces to 40×24 pixels.

The evaluation protocol for the YTF database is a set of ten-fold, cross validation, pair-matching tests. All subjects in each of the ten splits are "subject mutually exclusive" and include 250 same person (*i.e*., genuine) video pairs and 250 not-same (*i.e*., impostor) video pairs [24]. For all experiments in this study, we follow the *restricted* protocol. However, the COTS face matchers we evaluate are black box algorithms and may have been trained using data external to the YTF database. We report face verification results as mean and standard deviations of true accept rates (TAR) at fixed false accept rates (FAR) across all ten splits, and ROC curves are computed by threshold averaging.

Matchers: We evaluate the performance of the three COTS face matchers available to us. We have obfuscated the names and details of these specific matchers and instead refer to them as COTS-A, COTS-B, and COTS-C. All three matchers were participants in the 2010 NIST Multi-Biometric Evaluation (MBE) [8].

The accuracies of the proposed COTS fusion schemes are benchmarked against Wolf *et al*.'s Matched Background Similarity (MBGS) [24], Li *et al*.'s Adaptive Probabilistic Elastic Matching (APEM) Fusion [15], Cui *et al*.'s Spatio-Temporal Face Region Descriptor Pairwise-constrained Multiple Metric Learning (STFRD+PMML) [5], and Bhatt *et al*.'s method which we call Rank Aggregation [3]. The results of these methods, under the same protocol, are made publicly available[2] by the authors.

5.1. Experimental Results

We conduct three experiments for video-based face recognition using the YTF database. All experiments follow the 10-fold cross-validation pairwise tests protocol suggested for the YTF database [24].

Experiment 1: We first examine the performance of the three COTS matchers individually on the YTF video data. The match score between a pair of face tracks is obtained by multi-frame score-level fusion of the frame-to-frame scores of a COTS matcher as outlined in Sec. 4. The results for each of the three matchers with multi-frame fusion rules (mean, median, and max) are given in Table 2. We omit results for min fusion, as they were consistently inferior to the other rules. We report the average true accept rates (TAR) and standard deviations at fixed false accept rates (FAR) of 0.1% and 1.0%. Furthermore, to compare the performance

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

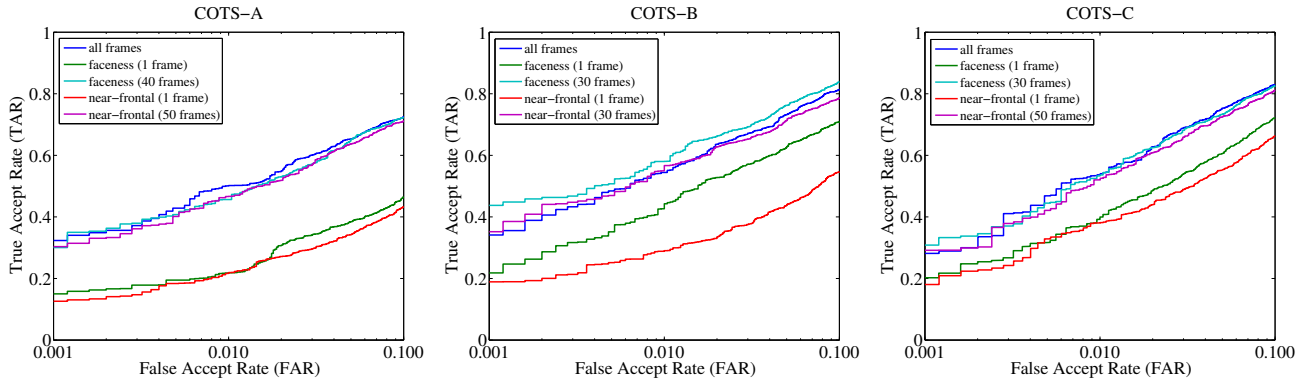

Figure 3. Verification results for key-frame selection based on two quality measures, highest faceness, and most near-frontal. Matching using the single highest quality frame from each track performs poorly for all COTS matchers; matching 30-50 frames from each track performs comparably to matching all frames for all matchers; matching the 30 frames with the highest faceness from each track improves the performance of the COTS-B matcher over matching all frames.

of a COTS matcher on video data versus still images, every frame-to-frame pair used in the video-to-video matching is considered as a verification attempt in isolation from their respective videos. This provides insight into how each COTS matcher performs on unconstrained still images.

For all three COTS matchers, the mean and median rules give the highest accuracies for video-to-video matching. These two methods consolidate all match scores obtained from comparing two face tracks and are thereby more representative of the entire tracks, whereas the min and max rules are susceptible to outliers in true and false matches, respectively. Because the YTF database is unconstrained, COTS matchers may output high similarity between an impostor pair or low similarity between a genuine pair of face images due to extreme variations in pose, expression, illumination, and/or resolution. In all subsequent experiments, we use the mean rule for multi-frame fusion.

MF Rule	COTS-A	COTS-B	COTS-C
mean	38.5 ± 7.9	40.4 ± 7.9	35.8 ± 9.6
median	36.8 ± 8.1	40.0 ± 6.5	34.7 ± 9.7
max	28.5 ± 5.1	22.6 ± 9.3	38.4 ± 5.9
frame-to-frame	25.2 ±15.3	37.2 ±11.9	38.7 ±12.5

(a) TAR @ FAR = 0.1%

MF Rule	COTS-A	COTS-B	COTS-C
mean	49.1 ± 3.2	55.6 ± 6.4	56.4 ± 5.8
median	47.6 ± 3.9	56.2 ± 6.3	55.2 ± 6.7
max	36.2 ± 4.8	40.8 ± 6.5	50.3 ± 4.2
frame-to-frame	36.0 ±14.7	55.0 ±11.8	56.6 ±11.9

(b) TAR @ FAR = 1.0%

Table 2. Verification results (with standard deviations) of three COTS matchers on the YTF database. Rows are multi-frame (MF) fusion rules for consolidating match scores obtained from all frame pairs of two video face tracks.

Table 2 also shows that the performance of COTS-A is significantly higher on video data than individual still images. COTS-B experiences a slight improvement on video data, and COTS-C performs comparably for video data and still images. For all three COTS matchers, the standard deviation of the TAR at a fixed FAR is reduced in the video-to-video matching for all score fusion rules. We believe this is because the multi-frame fusion rules suppress the effect of outliers on the overall accuracy for video matching.

Experiment 2: The results for quality-based key-frame selection are given in Fig. 3. To select a subset of key frames from each face track, we used each COTS matcher's own measure of faceness and frontal pose. We tried selecting sets of 1, 5, and 10-50 (in increments of 10) frames, but for sake of clarity have only included ROC curves for the best results with respect to the lowest number of frames. Overall, Fig. 3 demonstrates that COTS matchers are able to achieve consistent accuracies whether all frames or a subset of highest quality frames are matched from two face tracks. Furthermore, the use of more than one high quality frame for video-to-video matching improves the performance significantly. This shows that face recognition on multiple images from a video can help to achieve higher accuracy than on single unconstrained still images.

Note that for COTS-B, matching only the 30 frames with the highest faceness quality from each face track clearly improves performance over matching all frames of two face tracks, while for COTS-A and COTS-C, only a small improvement is obtained. Thus, the use of quality measures depends on the underlying COTS face matcher, as well as the accuracy of the quality measurement. However, this motivates the use of quality measures for improving the perfor-

MM Rule	tanh	z-score	min-max	median
sum	49.2 ±15.3	44.0 ± 8.8	44.4 ±10.5	45.2 ±16.1
median	48.8 ±15.3	46.2 ±11.9	29.5 ± 4.9	47.3 ±18.8
max	48.9 ±11.5	41.9 ± 8.8	41.1 ± 8.1	39.4 ±10.5

(a) TAR @ FAR = 0.1%

MM Rule	tanh	z-score	min-max	median
sum	63.1 ± 3.9	58.5 ± 5.7	58.8 ± 5.4	60.6 ± 7.3
median	62.9 ± 4.7	61.9 ± 3.7	43.4 ± 7.7	63.6 ± 6.7
max	59.8 ± 7.7	57.8 ± 6.8	56.4 ± 5.8	59.8 ± 5.3

(b) TAR @ FAR = 1.0%

Table 3. Verification results (with standard deviations) of MMMF fusion of three COTS matchers on YTF database. Columns are score normalization schemes; rows are multi-matcher (MM) fusion rules for the three COTS matchers operating on the same pair of frames.

mance of COTS static face matchers on video data.

Experiment 3: Lastly, we evaluate the performance of utilizing all three COTS matchers for video-to-video face recognition with MMMF and MFMM fusion, as outlined in Sec. 4.2. Parameters for all tanh, min-max, z-score, and median score normalization schemes are calculated using the scores from nine training splits when evaluating a test split. We calculated parameters for tanh normalization in the same manner as Jain *et al*. [10].

Results for the different score normalization schemes and multi-matcher fusion rules are given in Table 3 for MMMF. In total, we evaluated the performance of $4\times4\times4 = 64$ different combinations of score normalization schemes, multi-matcher fusion rules, and multi-frame fusion rules for both MMMF and MFMM pipelines. Due to space limitations, we omit most results for MFMM fusion as the performance was consistently comparable to MMMF. The best results for video-to-video face matching with individual COTS matchers, as well as fusion of multiple matchers, are shown in Fig. 4.

From Fig. 4 and Table 4, it is observed that all COTS matchers significantly outperform previous results on the YTF database, with all COTS matchers achieving roughly an order of magnitude decrease in error rates. Further, the fusion of all three COTS matchers achieves even greater accuracy, with a 20% accuracy improvement at low false accept rates. Perhaps equally important is that in using COTS matchers, the accuracies we demonstrate are through the use of systems that can be readily deployed in operational settings.

6. Conclusions

This paper has presented an evaluation of COTS static face matchers on video-to-video unconstrained face matching. Our results demonstrate that utilizing existing technology for video-based face recognition can achieve reasonable accuracies. COTS face matchers individually outperformed the best face recognition results previously published on the YTF database; further, a fusion of the three COTS matchers used here achieved a 20% improvement over previous results. We suggest subsequent research on video-to-video face matching should now demonstrate the ability to improve upon these presented baseline accuracies.

The increase in performance observed by fusion of multiple COTS face matchers can partially be attributed to the face registration problem (*e.g*. face detection and landmark localization). Table 5 gives the number of face images that each COTS matcher failed to enroll out of all the 587,035 frames in the YTF database. All three COTS matchers only failed to enroll the same 0.2% among all the frames, which is substantially lower than the number of frames that could not be enrolled in each COTS matcher individually. Thus, when we fuse the match scores obtained from multiple matchers, we are essentially also utilizing their complementary image registration capabilities. Fig. 5 shows examples of images from face tracks where one of the COTS matchers failed to enroll all frames of that track. This demonstrates that unconstrained face detection and landmark localization are crucial to be able to fully leverage all available frames in a face track.

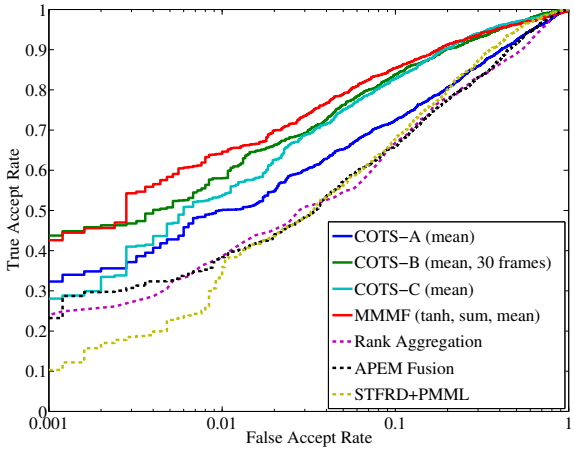

Figure 4. Verification results on the YTF database protocol. All three COTS face matchers and fusion of three matchers significantly outperform previous methods: Rank Aggregation [3], APEM Fusion [15], and STFRD+PMML [5].

Method	AUC	EER
MBGS [24]	82.6	25.3
APEM Fusion [15]	86.6	21.4
STFRD+PMML [5]	88.6	19.9
Rank Aggregation [3]	85.8	21.6

(a) Previous results

Method	AUC	EER
COTS-A	88.6	19.6
COTS-B	93.1	14.2
COTS-C	93.4	13.8
MMMF Fusion	93.9	12.6

(b) Our results

Table 4. AUC and EER statistics for performance comparisons on the YouTube Faces database.

COTS-A	COTS-B	COTS-C	COTS-All[a]	Total Frames
21,687	7,525	12,494	1,393	587,035

[a] All three COTS matchers failed to enroll these frames

Table 5. Number of frames that could not be enrolled out of the 587,035 total frames in the YTF database.

Figure 5. Example images from eight face tracks in the YTF database where all images in that track could not be enrolled by one of the COTS matchers. These images display extreme pose and illumination conditions, low resolution, and motion blur.

Acknowledgment

B. Klare's research was supported by the Noblis Sponsored Research (NSR) program. L. Best-Rowden's research was supported by the National Physical Science Consortium (NPSC) Graduate Fellowship and the National Security Agency (NSA).

References

[1] O. Arandjelovic and R. Cipolla. A manifold approach to face recognition from low quality video across illumination and pose using implicit super-resolution. *ICCV*, 1, 2007.

[2] J. H. Barr, K. W. Boyer, P. Flynn, and S. Biswas. Face recognition from video: A review. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 26(05), 2012.

[3] H. Bhatt, R. Singh, and M. Vatsu. On rank aggregation for face recognition from videos. In *IEEE ICIP*, 2013.

[4] S. Canavan, M. Kozak, Y. Zhang, J. Sullins, M. Shreve, and D. Goldgof. Face recognition by multi-frame fusion of rotating heads in videos. In *IEEE BTAS*, pages 1–6, 2007.

[5] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *IEEE CVPR*, 2013.

[6] R. Goh, L. Liu, X. Liu, and T. Chen. The CMU face in action (FIA) database. In *AMFG*, pages 255–263, 2005.

[7] R. Gross and J. Shi. The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Pittsburgh, PA, June 2001.

[8] P. J. Grother, G. W. Quinn, and P. J. Phillips. Report on the evaluation of 2D still-image face recognition algorithms. *NISTIR*, 7709, 2010.

[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. Technical Report 07-49, Univ. of Massachusetts, Amherst, October 2007.

[10] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285, 2005.

[11] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *IEEE CVPR*, pages 1–8, 2008.

[12] J. Kittler, J. Matas, K. Jonsson, and M. U. R. Sànchez. Combining evidence in personal identity verification systems. *Pattern Recognition Letters*, 18(9):845–852, 1997.

[13] K. Lee, J. Ho, M. Yang, and D. Kriegman. Visual tracking and recognition using probabilistic appearance manifolds. *CVIU*, 99(3):303–331, 2005.

[14] K.-C. Lee and D. Kriegman. Online learning of probabilistic appearance manifolds for video-based recognition and tracking. In *IEEE CVPR*, volume 1, pages 852–859, 2005.

[15] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *IEEE CVPR*, 2013.

[16] X. Liu and T. Chen. Video-based face recognition using adaptive hidden markov models. In *IEEE CVPR*, pages 340–345, 2003.

[17] U. Park, H. Chen, and A. K. Jain. 3d model-assisted face recognition in video. In *IEEE Canadian Conf. on Computer and Robot Vision*, pages 322–329, 2005.

[18] U. Park and A. K. Jain. Face recognition in video: Adaptive fusion of multiple matchers. In *IEEE CVPR*, pages 1–8, 2007.

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

[20] D. Thomas, K. W. Bowyer, and P. J. Flynn. Multi-frame approaches to improve face recognition. In *IEEE Workshop on Motion and Video Computing*, pages 19–19. IEEE, 2007.

[21] P. Viola and M. J. Jones. Robust real-time face detection. *Int. Journal of Computer Vision*, 57(2):137–154, 2004.

[22] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *IEEE CVPR*, pages 1–8, 2008.

[23] F. W. Wheeler, X. Liu, and P. H. Tu. Multi-frame super-resolution for face recognition. In *IEEE BTAS*, pages 1–6, 2007.

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

[25] M. Yang, P. Zhu, L. V. Gool, , and L. Zhang. Face recognition based on regularized nearest points between image sets. In *IEEE FG*, 2013.

[26] S. K. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Trans. on Image Processing*, 13(11):1491–1506, 2004.