Unconstrained Face Recognition: Identifying a Person of Interest From a Media Collection

Lacey Best-Rowden, *Student Member, IEEE*, Hu Han, *Member, IEEE*, Charles Otto, *Student Member, IEEE*, Brendan F. Klare, *Member, IEEE*, and Anil K. Jain, *Life Fellow, IEEE*

***Abstract*— As face recognition applications progress from constrained sensing and cooperative subjects scenarios (e.g., driver's license and passport photos) to unconstrained scenarios with uncooperative subjects (e.g., video surveillance), new challenges are encountered. These challenges are due to variations in ambient illumination, image resolution, background clutter, facial pose, expression, and occlusion. In forensic investigations where the goal is to identify a person of interest, often based on low quality face images and videos, we need to utilize whatever source of information is available about the person. This could include one or more video tracks, multiple still images captured by bystanders (using, for example, their mobile phones), 3-D face models constructed from image(s) and video(s), and verbal descriptions of the subject provided by witnesses. These verbal descriptions can be used to generate a face sketch and provide ancillary information about the person of interest (e.g., gender, race, and age). While traditional face matching methods generally take a single media (i.e., a still face image, video track, or face sketch) as input, this paper considers using the entire gamut of media as a probe to generate a single candidate list for the person of interest. We show that the proposed approach boosts the likelihood of correctly identifying the person of interest through the use of different fusion schemes, 3-D face models, and incorporation of quality measures for fusion and video frame selection.**

***Index Terms*— Unconstrained face recognition, uncooperative subjects, media collection, quality-based fusion, still face image, video track, 3D face model, face sketch, demographics.**

I. Introduction

As FACE recognition applications progress from constrained imaging and cooperative subjects (*e.g.,* identity card deduplication) to unconstrained imaging scenarios with uncooperative subjects (*e.g.,* watch list monitoring), a lack of guidance exists with respect to optimal approaches for integrating face recognition algorithms into large-scale applications of interest. In this work we explore the problem of identifying a person of interest given a variety of information sources about the person (face image, surveillance video, face sketch, 3D face model, and demographic information) in both closed set and open set identification modes.

Manuscript received March 15, 2014; revised July 2, 2014; accepted August 26, 2014. Date of publication September 19, 2014; date of current version November 10, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gérard Medioni.

L. Best-Rowden, H. Han, C. Otto, and A. K. Jain are with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824 USA (e-mail: bestrow1@msu.edu; hhan@msu.edu; ottochar@msu.edu; jain@msu.edu).

B. F. Klare is with Noblis, Falls Church, VA 22042 USA (e-mail: brendan.klare@noblis.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2014.2359577

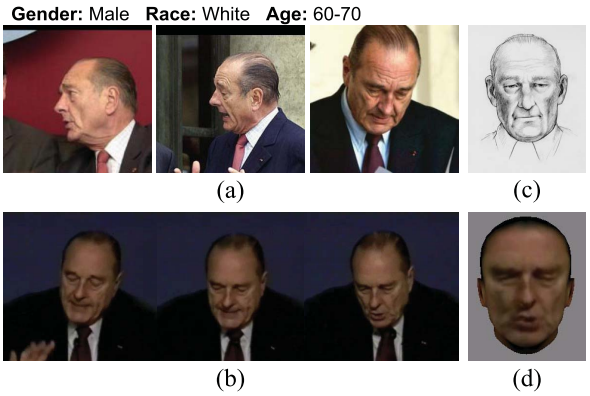


Fig. 1. A collection of face media for a particular subject may consist of (a) multiple still images, (b) a face track from a video, (c) a forensic sketch, (d) a 3D face model of the subject derived from (a) and/or (b), and demographic information (*e.g.,* gender, race, and age). The images and video track shown here are from [4] and [5]. The sketch was drawn by a forensic sketch artist after viewing the face video. In other applications, sketches could be drawn by an artist based on verbal description of the person of interest.

Identifying a person based on unconstrained face images is an increasingly prevalent task for law enforcement and intelligence agencies. In general, these applications seek to determine the identity of a subject based on one or more probe images or videos, where a top 200 ranked list retrieved from the gallery (for example) may suffice for analysts (or forensic examiners) to identify the subject [1]. In many cases, such a forensic identification is performed when multiple face images and/or a face track (*i.e.,* a sequence of cropped face images which can be assumed to be of the same person) from a video of a person of interest are available (see Fig. 1). For example, in investigative scenarios, multiple face images of an unknown subject often arise from an initial clustering of visual evidence, such as a network of surveillance cameras, the contents of a seized hard drive, or from open source intelligence (*e.g.,* social networks). In turn, these probe images are searched against large-scale face repositories, such as mug shot or identity card databases.

High profile crimes such as the Boston Marathon bombings often rely on data extracted by significant manual effort to identify the person of interest:

> "It's our intention to go through every frame of every video [from the marathon bombings]," Boston Police Commissioner Ed Davis[1]

[1] http://www.washingtonpost.com/world/national-security/boston-marathon-bombings-investigators-sifting-through-images-debris-for-clues/2013/04/16/1cabb4d4-a6c4-11e2-b029-8fb7e977ef71_story.html

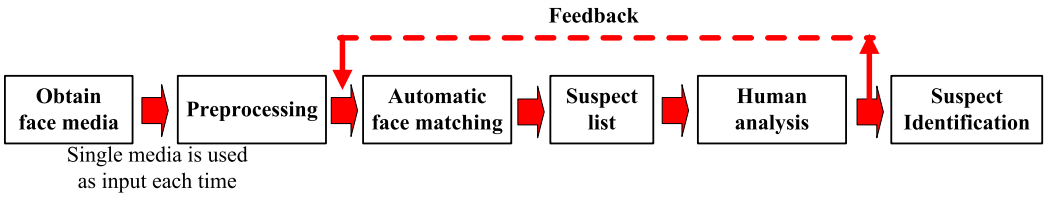


Fig. 2. Forensic investigations by law enforcement agencies using face images typically involve six main stages: obtaining face media, preprocessing, automatic face matching, generating a suspect list, human analysis, and suspect identification. Feedback occurs after human analysis reveals that, for example, additional preprocessing of the input image (*e.g.,* illumination correction and/or manual eye locations), demographic filtering of the gallery, and/or a different face sample from the media collection is necessary.

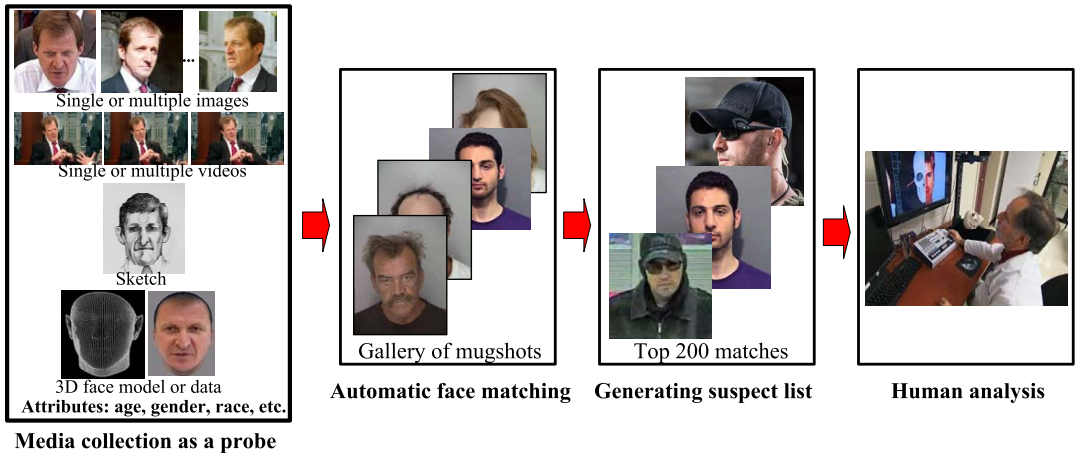


Fig. 3. Schematic diagram of a person identification task given a face media collection as input.

While other routine, but high value, crimes such as armed robberies, kidnappings, and acts of violence require similar identifications, only a fraction of the manual resources are available to solve these crimes. Thus, it is paramount for face recognition researchers and practitioners to have a firm understanding of optimal strategies for combining multiple sources of face information, collectively called *face media*, available to identify the person of interest.

While forensic identification is focused on human-driven queries, several emerging applications of face recognition technology exist where it is neither practical nor economical for a human to have a high degree of intervention with the automatic face recognition system. One such example is watch list identification from surveillance cameras, where a list of persons of interest are continuously searched against streaming videos. Termed as *open set recognition*, these challenging applications will likely have better success as unconstrained face recognition algorithms continue to develop and mature [2]. While a closed-set identification system deals with the scenario where the person of interest is assumed to be present in the gallery, and always returns a non-empty candidate list, an open-set identification system allows for the scenario where the person of interest is not enrolled in the gallery, and so can return a possibly empty candidate list [3]. We provide experimental protocols, recognition accuracies on these protocols using COTS face recognition and 3D face modeling algorithms, and an analysis of the integration strategies to improve operational scenarios involving open set recognition.

A. Overview

In forensic investigations, manual examination of a suspect's face image against a mug shot database with millions of face images is prohibitive. Thus, automatic face recognition techniques are utilized to generate a candidate suspect list. As shown in Fig. 2, forensic investigations using face images typically involve six stages: obtaining face media, preprocessing, automatic face matching, generating a suspect list, human or forensic analysis, and suspect identification.[2] The available forensic data or media of the suspect may include still face image(s), video track(s), a face sketch, and demographic information (*e.g.,* age, gender, and race) as shown in Fig. 3. While traditional face matching methods take a single media (*i.e.,* a still face image, video track, or face sketch) as probe to generate a suspect list, a media collection is expected to provide more identifiable information about a suspect. The proposed approach contributes to forensic investigations by taking into account the entire media collection of the suspect to perform face matching. This approach generates a single candidate suspect list (rather than a separate list for each face sample in the collection), thereby reducing the amount of human analysis needed.

In this paper, we examine the use of commercial off the shelf (COTS) face recognition systems with respect to the aforementioned challenges in large-scale unconstrained face recognition scenarios. First, the efficacy of forensic

[2]A more detailed description of this forensic investigation process can be found at: http://www.justice.gov/criminal/cybercrime/docs/forensics_chart.pdf

identification is explored by combining two public-domain unconstrained face databases, Labeled Faces in the Wild (LFW) [4] and YouTube Faces (YTF) [5], to create sets of multiple probe images and videos to be matched against a gallery consisting of a single image for each subject. To replicate forensic identification scenarios, we further populate our gallery with one million operational mug shot images from the Pinellas County Sheriff's Office (PCSO) database.[3] Using this data, we are able to examine how to boost the likelihood of face identification through different fusion schemes, incorporation of 3D face models and hand drawn sketches, and methods for selecting the highest quality video frames. Researchers interested in improving forensic identification accuracy can use this competitive baseline (on public-domain databases LFW and YTF) to provide more objectivity towards such goals.

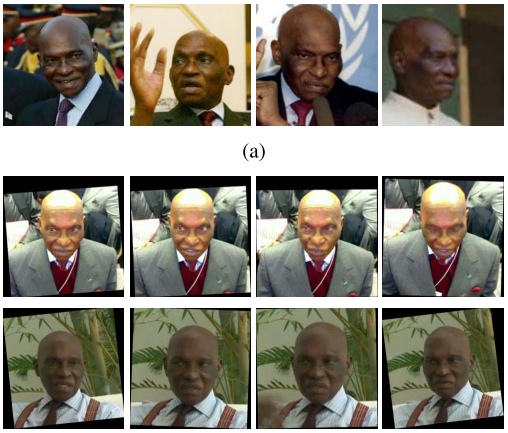
Most of the work on unconstrained face recognition using the LFW and YTF databases has been reported in verification scenarios [6], [7]. However, in forensic investigations, it is the identification mode that is of interest, especially the open-set identification scenario where the person of interest may not be present in legacy face databases.

The contributions of this work are summarized as follows:

- We show, for the first time, how a collection of face media (image(s), video(s), 3D model(s), demographic data, and sketch) can be used to mitigate the challenges associated with unconstrained face recognition (uncooperative subjects, unconstrained imaging conditions) and boost recognition accuracy.
- Unlike previous studies that report results in verification mode, we present results for both open set and closed set identifications which are the norm in identifying persons of interest in forensic and watch list scenarios.
- We present effective face quality measures to determine when the fusion of information sources will help boost identification accuracy. The quality measures are also used to assign weights to different media sources in fusion schemes.
- To demonstrate the effectiveness of media-as-input for the difficult problem of unconstrained face recognition, we utilize a state of the art COTS face matcher and a separate COTS 3D face modeler, namely the Aureus 3D SDK provided by CyberExtruder. Face sketches were drawn by forensic sketch artists who generated the sketch after viewing low quality videos. In the absence of demographic data for LFW and YTF databases, we used crowdsourcing to obtain the estimates of gender and race. The above strategy allows us to show the contribution of various media components as we incrementally add them as input to the face matching system.
- Pose-corrected versions of all face images in the LFW database, pose-corrected video frames from the YTF database, forensic sketches, and experimental protocols used in this paper have been made publicly available.[4]

[3]http://biometrics.org/bc2010/presentations/DHS/mccallum-DHS-Future-Opportunities.pdf

[4]http://biometrics.cse.msu.edu/pubs/databases.html



(a)

(b)

Fig. 4. Example (a) face images from the LFW database and (b) face video tracks from the YTF database. All faces shown are of the same subject.

The remainder of the paper is organized as follows. In Section II, we briefly review published methods related to unconstrained face recognition. We detail the proposed face media collection as input and media fusion method in Sections III and IV, respectively. Experimental setup and protocols are given in Section V, and experimental results are presented in Section VI. We conclude this work in Section VII.

II. Related Work

The release of the public-domain database Labeled Faces in the Wild[5] (LFW) in 2007 spurred interest and progress in unconstrained face recognition. The LFW database is a collection of 13,233 face images, downloaded from the Internet, of 5,749 different individuals such as celebrities, public figures, etc. [4]. These images were selected since they meet the criterion that faces can be successfully detected by the Viola-Jones face detector [8]. Despite this property, the LFW database contains significant variations in facial pose, illumination, and expression, and many of the face images are occluded. The LFW protocol consists of face verification based on ten-fold cross-validation, each fold containing 300 "same face" and 300 "not-same face" image pairs.

The YouTube Faces[6] (YTF) database, released in 2011, is the video-equivalent to LFW for unconstrained face matching in videos. The YTF database contains 3,425 videos of 1,595 individuals. The individuals in the YTF database are a subset of those in the LFW database. Faces in the YTF database were also detected with the Viola-Jones face detector at 24 fps, and face tracks were included in the database if there were at least 48 consecutive frames of that individual's face. Similar to the LFW protocol, the YTF face verification protocol consists of ten-fold cross-validation, each fold containing 250 "same face" and 250 "not-same face" track pairs. Figure 4 shows example face images and video tracks from the LFW and YTF databases for one particular subject. In this paper, we combine these two databases to evaluate the performance of face recognition on unconstrained face media collections.

[5]http://vis-www.cs.umass.edu/lfw/

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

TABLE I
A SUMMARY OF PUBLISHED METHODS ON UNCONSTRAINED FACE RECOGNITION (UFR). PERFORMANCE IS REPORTED AS TRUE ACCEPT RATE (TAR) AT A FIXED FALSE ACCEPT RATE (FAR) OF 0.1% OR 1%, UNLESS OTHERWISE NOTED

	Dataset	Scenario (query (size) vs. target (size))	Accuracy (TAR @ FAR)	Source
Single Media Based UFR	FRGC v2.0 Exp. 4 unconstrained vs. constrained	Single image (8,014) vs. single image (16,028)	12% @ 0.1%	Phillips *et al.* [9]
	MBGC v2.0 unconstrained vs. unconstrained	Single image (10,687) vs. single image (8,014)	97% @ 0.1%	Phillips *et al.* [9]
	MBGC v2.0 non-frontal vs. frontal	Single image (3,097) vs. single image (16,028)	17% @ 0.1%	Phillips *et al.* [9]
	MBGC v2.0 unconstrained vs. HD video	Single image (1,785) vs. single HD video (512)	94% @ 0.1%	Phillips *et al.* [9]
	MBGC v2.0 walking vs. walking	Notre Dame: Single video (976) vs. single video (976) UT Dallas: Single video (487) vs. single video (487)	Notre Dame: 46% @ 0.1% UT Dallas: 65% @ 0.1%	Phillips *et al.* [9]
	FRGC v2.0 Exp. 3 3D vs. 3D	Single 3D image (4,007) vs. single 3D image (4,007)	53% @ 0.1%	Phillips *et al.* [9]
	LFW Image-Restricted (Strict[3])	300 genuine and 300 impostor pairs per fold[1]	61% @ 1%[2]	Simonyan *et al.* [10]
	LFW Image-Unrestricted (Outside[3])	300 genuine and 300 impostor pairs per fold[1]	88% @ 1%[2]	Chen *et al.* [11]
	LFW Image-Unrestricted (Outside[3])	300 genuine and 300 impostor pairs per fold[1]	94% @ 1%[2]	Taigman *et al.* [12]
	LFW Image-Unrestricted (Outside[3])	300 genuine and 300 impostor pairs per fold[1]	95% @ 1%[2]	Sun *et al.* [13]
	LFW	4,249 subjects and 9,708 images per fold	42% @ 0.1% 66% @ 1%[4]	Liao *et al.* [14]
	YouTube Celebrities	1,500 video clips of 35 celebrities	Rank-1 acc.: 71%	Kim *et al.* [15]
	YouTube Faces	250 genuine and 250 impostor pairs per fold[1]	55% @ 1%[2]	Taigman *et al.* [12]
	YouTube Faces	250 genuine and 250 impostor pairs per fold[1]	63% @ 1%[2]	Best-Rowden *et al.* [16]
Media Collection Based UFR	FRGC v2.0 Exp. 3	Single image & single 3D image (8,014) vs. single 3D image (943)	79% @ 0.1%	Phillips *et al.* [9]
	MBGC v2.0 unconstrained face & iris vs. NIR & HD videos	Single image & single iris (14,115) vs. single NIR & single HD (562)	97% @ 0.1%	Phillips *et al.* [9]
	LFW	Single image vs. single image	56.7%	this paper[5]
		Multi-images vs. single image	72.0%	
	YouTube Faces	Single video vs. single image	31.3%	
		Multi-videos vs. single image	44.0%	
	3D face model	Multi-images & multi-videos vs. single image	77.5%	
	Forensic sketch	Multi-images, multi-videos, & 3D model vs. single image	83.0%	
	Demographic information	Multi-images, multi-videos, 3D model & demographics vs. single image	84.9%	

[1]Performance is an average across 10 folds. [2]About 40 different methods (e.g., [17]–[20]) have reported performance on LFW, but all of them can be classified as single media (image vs. image) based UFR methods. Due to limited space, we only list the most recently reported performance for each testing protocol in this table. Similarly, methods that have reported results on YTF are also single media (video vs. video) based UFR method. [3]Strict vs. outside: no outside training data is used vs. outside training data is used. [4]Performance is reported as mean minus standard deviation over 10 trials. [5]The performance of the proposed method is the Rank-1 identification accuracy.

We provide a summary of related work on unconstrained face recognition, focusing on various face media matching scenarios in Table I. We emphasize that most prior work has evaluated unconstrained face recognition methods in the verification mode. While fully automated face recognition systems are able to achieve ∼99% True Accept Rate (TAR) at 0.1% False Accept Rate (FAR) in constrained imagery and cooperative subject conditions, face recognition in unconstrained environments remains a challenging problem [9]. However, face verification accuracies on the LFW protocol have recently seen drastic improvements. When utilizing outside training data, recent works have achieved TARs greater than 94% at 1% FAR and classification accuracies over 97% (see [12], [13]). However, the LFW protocol only contains three impostor scores at 1% FAR, so these saturated accuracies may overestimate the abilities of FR systems on unconstrained faces. Liao *et al.* propose a new benchmark for LFW which allows for evaluation at lower FARs; out of three features

and seven learning algorithms, they find the best performance is 42% and 66% at 0.1% and 1% FAR, respectively [14]. Open-set identification performance is even lower at 18% for Rank-1 and 1% FAR [14].

Unconstrained face recognition methods can be grouped into two main categories: single face media based methods and face media collection based methods. Single media based methods focus on the scenario where both the query and target instances contain only one type of face media, such as a still image(s), video track(s), or 3D image(s) or model(s). However, the query and target instances can be different media types, such as single image vs. single video. These methods can be effective for unconstrained illumination and expression variations but can only handle limited pose variations. For example, while ∼97% TAR at 0.1% FAR has been reported in MBGCv2.0 unconstrained vs. unconstrained face matching, under large pose variations, this performance drops to ∼17% TAR in MBGCv2.0 non-frontal vs. frontal face matching (see Table I). Such challenges were also observed in single image vs. single image face matching in LFW, and single video vs. single video face matching in YTF and MBGCv2.0 walking vs. walking databases.

These observations suggest that in unconstrained scenarios, a single face media probe, especially of "low quality", may not be able to provide a sufficient description of a face. This motivates the use of a face media collection which utilizes any source of information that is available for a probe (or query) instance of a face. One preliminary study in this direction is the FRGCv2.0 Exp. 3 where (i) a single 3D face image and (ii) a collection of single 3D image and a single 2D face image were used as queries. Results show that 2D face image and 3D face image did improve the face matching performance (79% TAR for 3D face and 2D face vs. 53% TAR for just the 3D face at 0.1% FAR) in unconstrained conditions. It is, therefore, important to determine how we can improve the face matching accuracy when presented with a collection of face media of different types, albeit of different qualities, as probe.

III. Media-as-Input

A face media collection can consist of still images, video tracks, a 3D model, a forensic sketch, and demographic information. In this section, we discuss how we use face "media-as-input" as probe and our approach to media fusion.

A. Still Image and Video Track

Still image and video track are two of the most widely used sources of media in face recognition systems [3]. Given multiple still images and videos, we use the method reported in [16] to match all still images and video frames available for a subject of interest to the gallery mugshot (frontal pose) images using a COTS face matcher. The resulting match scores are then fused to get a single match score for either multiple probe images or video(s).

B. 3D Face Models

One of the main challenges in unconstrained face recognition is large variations in facial pose [21], [22]. In particular, out-of-plane rotations drastically change the 2D appearance of a face, as they cause portions of the face to be occluded. A common approach to mitigate the effects of pose variations is to build a 3D face model from a 2D image(s) so that synthetic 2D face images can then be rendered at designated poses (see [23]–[25]).

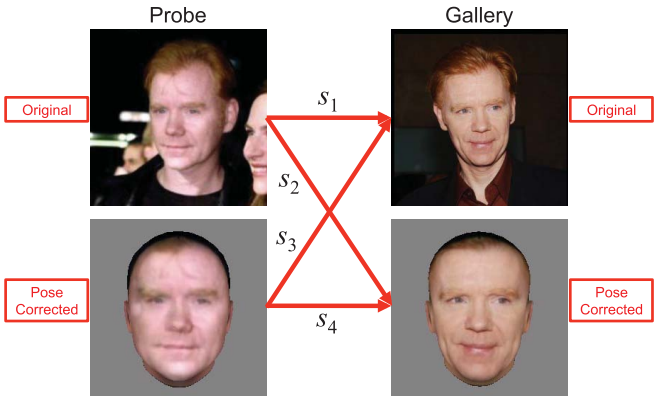


Fig. 5. Pose correction of probe (left) and gallery (right) face images using CyberExtruder's Aureus 3D SDK. We consider the fusion of four different match scores (s_1, s_2, s_3, and s_4) between the original probe and gallery images (top) and synthetic pose corrected probe and gallery images (bottom).

In this paper, we use a state of the art COTS 3D face modeling SDK, namely CyberExtruder's Aureus 3D SDK, to build 3D models from 2D unconstrained face images.[7] We input eye locations (extracted automatically by [11] for LFW images and the COTS face matcher for YTF video frames) to the SDK to help with model robustness. The entire 3D face modeling process is fully automatic. The 3D face model is then used to render a "pose corrected" (*i.e.,* frontal facing) image of the unconstrained probe face images. The pose corrected image can then be matched against a frontal gallery. We also pose correct "frontal" gallery images because even the gallery images can have variations in pose as well. Experimental results show that including pose corrected gallery images indeed improves the identification performance.

Given the original and pose corrected probe and gallery images, there are four matching scores that can be computed between any pair of probe and gallery face images (see Fig. 5). We use the score s_1 as the baseline to determine whether including scores s_2, s_3, s_4, or their fusion can improve the performance of a COTS face matcher. A face in a video frame can be pose corrected in the same manner. The Aureus SDK also summarizes faces from multiple frames in a video track as a "consolidated" 3D face model (see Fig. 6).

C. Demographic Attributes

In many law enforcement and government applications, it is customary to collect ancillary information like age, gender, race, height, and eye color from the subjects during enrollment. We explore how to best utilize demographic data to boost the recognition accuracy. Demographic information such as age, gender and race becomes even more important in complementing identity information provided by face images and videos

[7] http://www.cyberextruder.com/aureus-3d-sdk

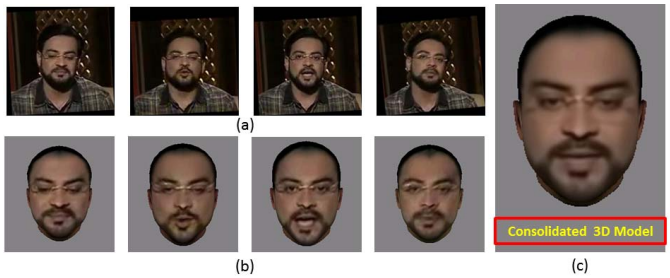


Fig. 6. Pose corrected faces (b) in a video track (a) and the resulting "consolidated" 3D face model (c). The consolidated 3D face model is a summarization of all frames in the video track.

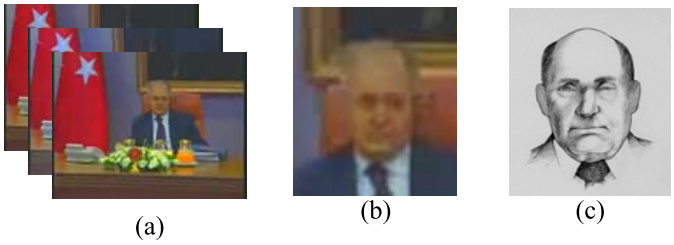


Fig. 7. An example of a sketch drawn by a forensic artist by looking at a low-quality video. (a) Video shown to the forensic artists, (b) facial region cropped from the video frames, and (c) sketch drawn by the forensic artist. Here, no verbal description of the person of interest is available.

in unconstrained face recognition due to the difficulty of the face matching task.

In this paper, we take gender and race attributes of each subject in the LFW and YTF face databases as one type of media. Since this demographic information is not available for the subjects in the LFW and YTF face databases, we utilized the Amazon Mechanical Turk (MTurk) crowdsourcing service[8] to obtain the "ground-truth" gender, and race of the 596 subjects that are common in LFW and YTF datasets. Most studies on automatic demographic estimation are limited to frontal face images [26]; demographic estimation from unconstrained face images (*e.g.,* the LFW database) is challenging [27]. For gender and race estimation tasks, we submitted 5, 749 (*i.e.,* the number of subjects in LFW) Human Intelligence Tasks (HITs), with ten human workers per HIT, at a cost of 2 cents per HIT. Finally, a majority voting scheme (among the responses) was utilized to determine the gender (Female or Male) and race (Black, White, Asian or Unknown) of each subject. We did not consider age in this paper due to large variations in age estimates by crowd workers.

D. Forensic Sketches

Face sketch based identification dates back to the 19th century [28], where the paradigm for identifying subjects using face sketches relied on human examination. Recent studies on automated sketch based identification systems show that sketches can also be helpful to law-enforcement agencies to identify the person of interest from mugshot databases [29], [30]. In situations where the suspect's photo or video is not available, expertise of forensic sketch artists are utilized to draw a suspect's sketch based on a verbal description provided by an eyewitness or victim. In some situations, even when a photo or video of a suspect is available, the quality of this media can be poor. In this situation also, a forensic sketch artist can be called in to draw a face sketch based on the low-quality face photo or video. For this reason, we also include the face sketch in a face media collection.

We manually selected 21 low-quality (large pose variations, shadow, blur, etc.) videos (one video per subject) from the YTF database (for three subjects, we also included a low quality still image from LFW). We then asked two forensic sketch artists to draw a face sketch for each subject in these videos (10 subjects were drawn by one forensic sketch artist, and 11 subjects by the other). Our current experiments are limited to sketches of 21 subjects due to the high cost of hiring a sketch artist. Examples of these sketches and their corresponding low-quality videos are shown in Figs. 7 and 15.

[8]www.mturk.com/mturk/

IV. Media Fusion

Given a face media collection as probe, there are various schemes to integrate the identity information provided by each individual media component, such as score level, rank level, and decision level fusion [31]. Among these approaches, score level fusion is the most commonly adopted. Some COTS matchers do not output a meaningful match score (to prevent hill-climbing attacks [32]). Thus, in these situations, rank level or decision level fusion is typically adopted.

In this paper, we match each face media (image, video, 3D model, sketch, or demographic information) of a probe collection to the gallery and combine the scores using score level fusion. Specifically, score level fusion takes place in two different layers: (i) fusion within one type of media, and (ii) fusion across different types of media. The first fusion layer generates a single score from each media type if multiple instances are available. For example, matching scores from multiple images or multiple video frames can be fused to get a single score. Additionally, if multiple video clips are available, matching scores of individual video clips can also be fused. Score fusion within the ith face media can generally be formulated as

$$s_i = \mathfrak{F}(s_{i,1}, s_{i,2}, \ldots, s_{i,n}), \tag{1}$$

where s_i is a single match score based on n instances of the ith face media type; $\mathfrak{F}(\cdot)$ is a score level fusion rule; we use the *sum* rule, *e.g.,* $s = \frac{1}{n}\sum s_{i,n}$, which has been found to be quite effective in practice [16]. Note that the *sum* and *mean* rules are equivalent, but we use the terms *mean* and *sum* for situations when normalization by the number of scores is and is not necessary, respectively. Given a match score for each face media type, the next fusion step involves fusing the scores across different types of face media. Again, the *sum* rule is used and found to work very well in our experiments; however, as shown in Fig. 8, face media for a person of interest can be of different quality. For example, a 3D face model can be corrupted due to inaccurate localization of facial landmarks. As a result, match scores calculated from individual media sources may have different degrees of confidence.

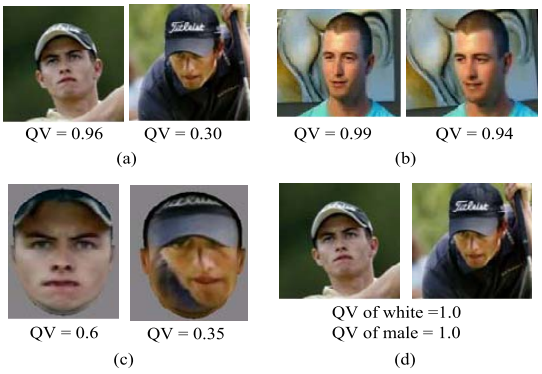


Fig. 8. Examples of different face media types with varying quality values (QV) of one subject: (a) images, (b) video frames, (c) 3D face models, and (d) demographic information. The range of QV is [0, 1].

We take into account the quality of individual media type by designing a quality based fusion. Specifically, let $\mathbf{S} = [s_1, s_2, \ldots, s_m]^T$ be a vector of the match scores between n different media types in a collection of probe and gallery, and $\mathbf{Q} = [q_1, q_2, \ldots, q_m]^T$ be a vector of quality values for the corresponding input media. Match scores from the COTS matcher are normalized with *z-score* normalization. The quality values are normalized to the range [0, 1]. The final match score between a probe and a gallery image is calculated by a weighted *sum* rule fusion,

$$s = \frac{1}{m}\sum_{i=1}^{m} q_i s_i = \mathbf{Q}^T\mathbf{S}. \tag{2}$$

Note that the quality based across-media fusion in (2) can also be applied to score level fusion within a particular face media type (*e.g.,* 2D video frames).

In this paper, we have considered five types of media in a collection: 2D face image, video, 3D face model, sketch, and demographic information. However, since sketches of only 21 persons (out of 596 persons that are common in LFW and YTF databases) are available, in most of the experiments, we perform quality-based fusion in (2) based on only four types of media ($m = 4$). The quality measures for individual media type are defined as follows.

- **Image and video:** For a probe image, the COTS matcher assigns a face confidence value in the range of [0, 1], which is used as the quality value. For each video frame, the same face confidence value measure is used. The average face confidence value across all frames is used as the quality value for a video track.
- **3D face model:** The Aureus 3D SDK used to build a 3D face model from image(s) or video frame(s) does not output a confidence score. We define the quality of a 3D face model based on the pose corrected 2D face image generated from it. Given a pose corrected face image, we calculate its structural similarity (SSIM) [33] to a set of predefined reference images (manually selected frontal face images). Let $\mathbf{I}_{PC}$ be a pose corrected face image (from the 3D model), and $\mathbf{R} = \{\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_t\}$ be the set of t reference face images. The quality value of a 3D model based on SSIM is defined as

$$\begin{aligned} q(\mathbf{I}_{PC}) &= \frac{1}{t}\sum_{i=1}^{t} \mathrm{SSIM}(\mathbf{I}_{PC}, \mathbf{R}_i) \\ &= \frac{1}{t}\sum_{i=1}^{t} l(\mathbf{I}_{PC}, \mathbf{R}_i)^{\alpha} \cdot c(\mathbf{I}_{PC}, \mathbf{R}_i)^{\beta} \cdot s(\mathbf{I}_{PC}, \mathbf{R}_i)^{\gamma} \end{aligned} \tag{3}$$

  where $l(\cdot)$, $c(\cdot)$, and $s(\cdot)$ are luminance, contrast, and structure comparison functions [33], respectively; α, β, and γ are parameters used to adjust the relative importance of the three components. We use the recommended parameters $\alpha = \beta = \gamma = 1$ in [33]. The quality value is in the range of [0, 1].
- **Demographic information:** As stated earlier, we collected demographic attributes (gender and race) of each face image using the MTurk crowdsourcing service with ten MTurk workers per task. Hence, the quality of demographic information can be measured by the degree of consistency among the ten MTurk workers. Let $\mathbf{E} = [e_1, e_2, \cdots, e_k]^T$ be the collection of estimates of one specific demographic attribute (gender or race) by k (here, $k = 10$) MTurk workers. The quality value of this demographic attribute can be calculated as

$$q(\mathbf{E}) = \frac{1}{k}\max_{i=1,2,\cdots,c}\{\sum(\mathbf{E} == i)\}, \tag{4}$$

  where c is the total number of classes for one demographic attribute. Here, $c = 2$ for gender (Male and Female); while $c = 4$ for race (Black, White, Asian, and Unknown). The notation $\sum(\mathbf{E} == i)$ denotes the number of estimates that are labeled as class i. The quality value range in (4) is in [0, 1].

Quality values for different face media of one subject are shown in Fig. 8. We note that the proposed quality measures give reasonable quality assessments for different input media.

V. Experimental Setup

The 596 subjects who have at least two images in the LFW database and at least one video track in the YTF database (subjects in YTF are a subset of those in LFW) are used to evaluate the performance of face identification on media-as-input in both closed-set and open-set scenarios. The state of the art COTS face matcher used in our experiments was one of the top performers in the 2010 NIST Multi-Biometric Evaluation [9]. Though the COTS face matcher is designed for matching still images, we apply it to video-to-still face matching via multi-frame fusion to obtain a single score for the video track [16]. In all cases where video tracks are part of the face media collection, we use the *mean* rule for multi-frame fusion (the *max* fusion rule performed comparably [16]).

A. Closed Set Identification

In closed set identification experiments, one frontal LFW image per subject is placed in the gallery (one with the highest frontal score from the COTS matcher), and the remaining LFW images are used as probes. All YTF video

TABLE II
NUMBER OF PROBE FACE IMAGES (FROM THE LFW DATABASE) AND VIDEO TRACKS (FROM THE YTF DATABASE) AVAILABLE FOR THE 596 SUBJECTS THAT ARE COMMON IN THE TWO DATABASES

# images/videos per subj.	1	2	3	4	5	6	7+
# subjects (LFW images)	238	110	78	57	25	12	76
# subjects (YTF videos)	204	190	122	60	18	2	0

tracks for the 596 subjects are used as probes. Table II shows the distribution of number of probe images and videos per subject. The average number of images, video tracks, and total media instances per subject is 5.3, 2.2, and 7.4, respectively. We further extend the gallery size with an additional 3,653 LFW images (of subjects with only a single image in LFW). In total, the size of the gallery is 4,249.

We evaluate five different scenarios depending on the contents of the probe set: (i) single image as probe, (ii) single video track as probe, (iii) multiple images as probe, (iv) multiple video tracks as probe, and (v) multiple images and video tracks as probe. We also take into account the 3D face models and demographic information in the five scenarios. To better simulate the scenarios in real-world forensic investigations, we also provide a case study on the Boston Marathon bomber to determine the efficacy of using media, and the generalization ability of our system to a large gallery with one million background face images.

For all closed set experiments involving still images from LFW, we input automatically extracted eye locations (from [11]) to the COTS face matcher to help with enrollment because the COTS matcher sometimes enrolls a background face in the LFW image that is not the subject of interest. Against a gallery of approximately 5,000 LFW frontal images, we observed a 2–3% increase in accuracy for Rank-20 and higher by inputting the automatically extracted eye locations from [11]. Note that for the YTF video tracks, there are no available ground-truth eye locations for faces in each frame. Recall from Section III-B that we input eye locations from [11] and the COTS face matcher to build the 3D models for LFW images and YTF video frames, respectively; hence, the entire 3D face modeling process is fully automatic. We report closed set identification results as Cumulative Match Characteristic (CMC) curves.

B. Open Set Identification

Here, we consider the case when the person of interest in the probe image or video track may not have a true mate in the gallery. This is representative of a watch list scenario. The gallery (watch list) consists of 596 subjects with at least two images in the LFW database and at least one video in the YTF database. To evaluate performance in the open set scenario, we construct two probe sets: (i) a *genuine probe set* that contains faces matching gallery subjects, and (ii) an *impostor probe set* that does not contain faces matching gallery subjects.

We conduct two separate experiments: (i) randomly select one LFW image per watch list subject as the genuine probe set and use the remaining LFW images of subjects not on the watchlist as the impostor probe set (596 gallery subjects, 596 genuine probe images, and 9,494 impostor probe images), and (ii) use one YTF video per watch list subject as the genuine probe set, and the remaining YTF videos which do not contain watch list subjects as the impostor probe set (596 gallery subjects, 596 genuine probe videos, and 2,064 impostor probe videos). For each of these experiments, we evaluate three scenarios for the gallery: (i) single image, (ii) multiple images, and (iii) multiple images and videos.

Open set identification can be considered a two step process: (i) decide whether or not to reject a probe image as not in the watchlist, and (ii) if probe is in the watchlist, recognize the person. Hence the performance is evaluated based on (i) Rank-1 detection and identification rate (DIR), which is the fraction of genuine probes matched correctly at Rank-1, and not rejected at a given threshold, and (ii) the false alarm rate (FAR) of the rejection step (i.e. the fraction of impostor probe images which are not rejected). We report the DIR vs. FAR curve describing the tradeoff between true Rank-1 identifications and false alarms.

VI. EXPERIMENTAL RESULTS

A. Pose Correction

We first investigate whether using a COTS 3D face modeling SDK to pose correct a 2D face image prior to matching improves the identification accuracy. The closed set experiments in this section consist of a gallery of 4,249 frontal LFW images and a probe set of 3,143 LFW images or 1,292 YTF videos. Table III(a) shows that the COTS face matcher performs better on face images that have been pose corrected using the Aureus 3D SDK. Matching the original gallery images to the pose corrected probe images (*i.e.,* match score s_3) performs the best out of all four match scores, achieving a 7.25% improvement in Rank-1 accuracy over the baseline (*i.e.,* match score s_1). Furthermore, fusion of all four scores (s_1, s_2, s_3, and s_4) with the simple *sum* rule provides an additional 2.6% improvement at Rank-1. Consistent with the results for still images, match scores s_3 and $sum(s_1, s_2, s_3, s_4)$ also provide significant increases in identification accuracy over using match score s_1 alone for matching frames of a video track (Table III(b)). We note that s_4 likely performs lower than s_3 because the gallery images are already fairly frontal. If both the gallery and the probe face images are unconstrained then s_4 may perform better.

Next, we investigate whether the Aureus SDK consolidated 3D models (*i.e., n* frames of a video track summarized as a single 3D face model rendered at frontal pose) can achieve comparable accuracy to matching all n frames. Table V(a) shows that the accuracy of $sum(s_3, s_4)$ (*i.e.,* consolidated 3D models matched to original and pose corrected gallery images) provides the same accuracy as matching all n original frames (*i.e.,* score s_1 in Table III(b)). However, the accuracy of the consolidated 3D model is slightly lower ($\sim$5%) than *mean* fusion over all n pose corrected frames (*i.e.,* score s_3 in Table III(b)). Hence, the consolidated 3D model built from a video track is not able to retain all discriminatory information contained in the collection of n pose-corrected frames.

TABLE III

CLOSED SET IDENTIFICATION ACCURACIES (%) FOR POSE CORRECTED GALLERY AND/OR PROBE FACE IMAGES USING 3D MODEL. THE GALLERY CONSISTS OF 4,249 LFW FRONTAL IMAGES AND THE PROBE SETS ARE (*a*) 3,143 LFW IMAGES AND (*b*) 1,292 YTF VIDEO TRACKS. PERFORMANCE IS SHOWN AS RANK RETRIEVAL RESULTS AT RANK-1, 20, 100, AND 200. COMPUTATION OF MATCH SCORES s_1, s_2, s_3, AND s_4 ARE SHOWN IN FIG. 5

LFW Images

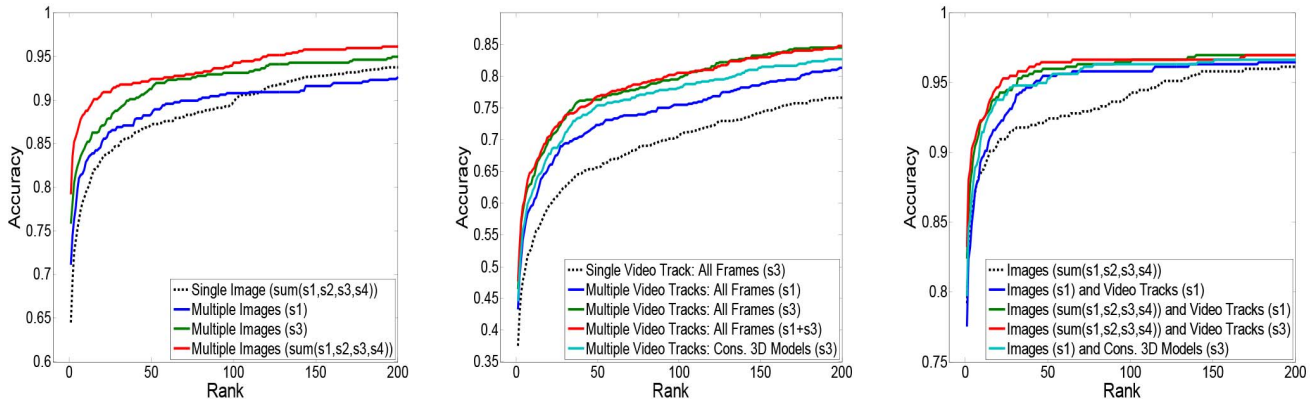
	R-1	R-20	R-100	R-200
s_1	56.7	78.1	87.1	90.2
s_2	57.7	77.6	86.0	89.9
s_3	**63.9**	**83.4**	**90.7**	**93.6**
s_4	55.6	78.8	88.0	91.9
sum	**66.5**	**85.9**	**92.4**	**95.1**

(a)

YTF Video Tracks

	R-1	R-20	R-100	R-200
s_1	31.3	54.2	68.0	74.5
s_2	32.3	55.3	67.8	73.9
s_3	**36.3**	**58.8**	**71.3**	**77.2**
s_4	31.7	54.4	68.7	76.5
sum	**38.8**	**61.4**	**73.6**	**79.0**

(b)



(a) Images (b) Video Tracks (c) Media Collection

Fig. 9. Closed set identification results for different probe sets: (a) multiple still face images, (b) multiple face video tracks, and (c) face media collection (images, videos and 3D face models). Single face image and video track results are plotted in (a) and (b) for comparison. Note that the ordinate scales are different in (a), (b), and (c) to accentuate the difference among the plots.

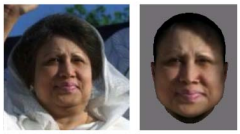
B. Forensic Identification: Media-as-Input

A summary of results for the various media-as-input scenarios is shown in Fig. 9. For all scenarios that involved multiple probe instances (*i.e.,* multiple images and/or videos), the *mean* fusion method gave the best result. For brevity, all CMC curves and results that involve multiple probe instances are also obtained via *mean* fusion. We also investigated the performance of rank-level fusion; the highest-rank fusion performed similar to score-level fusion, while the Borda count method [31] performed worse.

As observed in the previous section, pose correction with the Aureus 3D SDK to obtain scores s_3 or $sum(s_1, s_2, s_3, s_4)$ achieves better accuracies than score s_1. This is also observed in Figs. 9(a) and 9(b) where scores $sum(s_1, s_2, s_3, s_4)$ and s_3 provide approximately a 5% increase in accuracy over score s_1 for multiple images and multiple videos, respectively. This improvement is also observed in Fig. 9(c) for matching media that includes both still images and videos, but the improvement is mostly at low ranks (< Rank-50).

Figure 9 shows that (i) multiple probe images and multiple probe videos perform better than their single instance counterparts, but (ii) multiple probe videos actually perform worse than single probe image (see Figs. 9(a) and 9(b)). This is likely due in part to videos in the YTF database being of lower quality than the still images in the LFW database.

(a) Probe media collection (image, 3D model, and video track)



(b) Gallery true mate (image and 3D model)

Fig. 10. A collection of face media for a subject (a) consisting of a single still image, 3D model, and video track improves the retrieval rank of the true mate in the gallery (b). Against a gallery of 4,249 frontal images, the single still image was matched at Rank-438 with the true mate. Including the 3D model along with the still image improved the match to Rank-118, while the entire probe media collection was matched to the true mate at Rank-8.

However, we note that though multiple videos perform poorly compared to still images, there are still cases where the fusion of multiple videos with the still images does improve the identification performance. This is shown in Fig. 9(c); the best result for multiple images is plotted as a baseline to show that the addition of videos to the media collection improves identification accuracy. An example of this is shown in Fig. 10. For this particular subject, there is only a single

(a) Probe image and 3D model (b) Gallery image and 3D model

(c) Probe video tracks

Fig. 11. Having additional face media does not always improve the identification accuracy. In this example, the probe image with its 3D model (a) was matched at Rank-5 against a gallery of 4,249 frontal images. (b) Inclusion of three video tracks of the subject (c) to the probe set degraded the true match to Rank-216.

probe image available that exhibits extreme pose. The additional information provided by the 3D model and video track improves the true match from Rank-438 to Rank-8. In fact, the performance improvement of media (*i.e.,* multiple images and videos) over multiple images alone can mostly be attributed to cases where there is only a single probe image with large pose, illumination, and expression variations.

While Fig. 9 shows that including additional media to a probe collection improves identification accuracies on average, there are cases where matching the entire media collection can degrade the matching performance. An example is shown in Fig. 11. Due to the fairly low quality of the video tracks, the entire media collection for this subject is matched at Rank-216 against the gallery of 4, 249 images, while the single probe image and pose corrected image (from the 3D model) are matched at Rank-5. This necessitates the use of quality measures to assign a degree of confidence to each media.

We evaluated the face verification performance using the same database as the closed-set identification protocol (*i.e.,* gallery (target) of 4,249 images and probe (query) media collections of 596 subjects). We found that score s_3 still outperforms s_1, s_2, and s_4 for still images and videos frames. In investigating why s_3 performs better than s_4, we found that s_4 provides a better genuine score distribution than s_3, but the impostor distribution of s_4 has a longer tail. We believe this is partially due to similarities in the contours of two pose-corrected images. However, we find that multiple images with their 3D models (sum(s_1, s_2, s_3, s_4)) perform better than a media collection of multiple images (s_1) and video frames (s_1 or consolidated 3D model), whereas in closed-set identification, these media collections perform better than the multiple images and 3D models alone. In both identification and verification modes, the best performance is a collection of images with their 3D models and video frames (see Fig. 12). Image and video scores were normalized with *z-score* normalization.

C. Quality-Based Media Fusion

In this section, we evaluate the proposed quality measures and quality-based face media fusion. As discussed in Section IV, quality measures and quality-based face media fusion can be applied at both within-media layer and across-media layer.

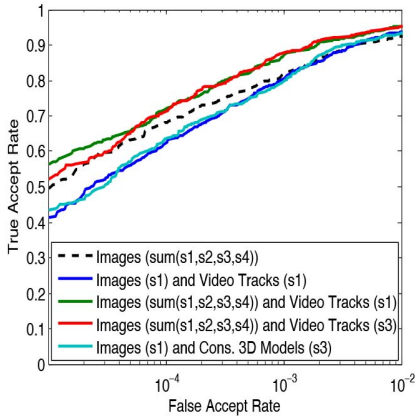


Fig. 12. Face verification performance of a gallery (target) of 4,249 frontal LFW images and probe (query) media collections of 596 subjects.

Tables IV(a) and (b) show the closed set identification accuracies of quality-based fusion of match scores ($s_1, \ldots, s_4$) of single image per probe and multiple images per probe, respectively. The performance with *sum* rule fusion is also provided for comparison. Our results indicate that the proposed quality measures and quality based fusion are able to improve the matching accuracies in both scenarios. Examples where the quality-based fusion performs better than *sum* rule fusion are shown in Fig. 13(a). Although in some cases the quality-based fusion may perform worse than *sum* rule fusion (see Fig. 13(b)), overall, it still improves the matching performance (see Table IV).

We have also applied the proposed quality measure for 3D face model to select high-quality frames that are used to build a consolidated 3D face model for a video clip. Figure 14(a) shows two examples where the consolidated 3D models using frame selection with SSIM quality measure (see Sec. IV) gets better retrieval ranks than using all frames. Although, a single value, *e.g.,* the SSIM based quality measure, may not always be reliable to describe the quality of a face image (see Fig. 14 (b)), frame selection still slightly improves the identification accuracy of the consolidated 3D face models at low ranks (see Table V).

D. Forensic Sketch Experiments

In this experiment, we study the effectiveness of forensic sketches in a media collection. For each subject with a forensic sketch, we input the forensic sketch to the COTS matcher to obtain a retrieval rank. Among the 21 subjects for whom we have a sketch, sketches of 12 subjects are observed to perform significantly better than the corresponding low-quality videos. Additionally, when demographic filtering using gender and race is applied, we can further improve the retrieval ranks. Figure 15 shows three examples where the face sketches significantly improved the retrieval ranks compared to low quality videos. The retrieval ranks of sketch and low-quality video fusion are also reported in Fig. 15.

To further demonstrate the efficacy of forensic sketch, we focus on identification of Tamerlan Tsarnaev, the older brother involved in the 2013 Boston Marathon bombing. In an earlier study Klontz and Jain [34] showed that while the younger

TABLE IV
CLOSED SET IDENTIFICATION ACCURACIES (%) FOR QUALITY BASED FUSION (QBF)
(*a*) WITHIN A SINGLE IMAGE, AND (*b*) ACROSS MULTIPLE IMAGES

QBF within a single image

	R-1	R-20	R-100	R-200
sum	65.7	83.2	90.1	93.5
QBF	66.5	85.9	92.6	95.3

(a)

QBF across multiple images

	R-1	R-20	R-100	R-200
sum	79.4	91.1	94.5	96.5
QBF	80.0	91.8	94.5	96.5

(b)

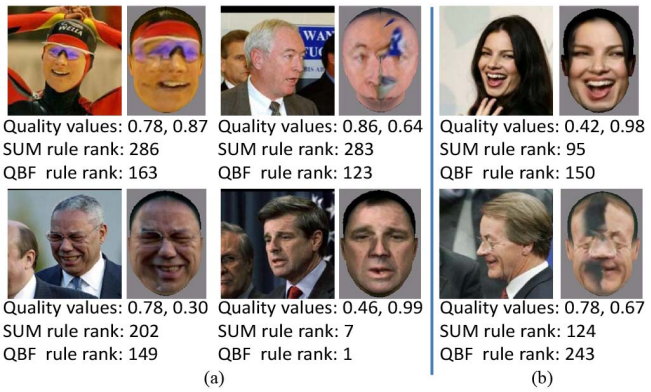


Fig. 13. A comparison of quality based fusion (QBF) vs. simple *sum* rule fusion (SUM). (a) Examples where quality based fusion provides better identification accuracy than *sum* fusion. (b) Examples where quality based fusion leads to lower identification accuracy compared with *sum* fusion.

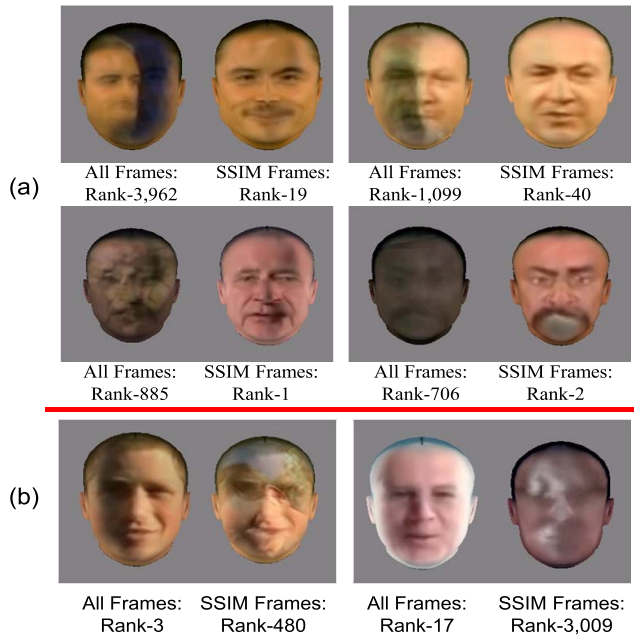


Fig. 14. Retrieval ranks using consolidated 3D face models (built from video tracks). Frame selection with SSIM quality measure (see Sec. IV) prior to building the consolidated 3D face model (a) improves and (b) degrades the identification accuracy. However, overall, frame selection using the proposed quality measure based on SSIM improves the COTS matcher's performance by an average of 1.43% for low ranks 1 to 50.

brother, Dzhokhar Tsarnaev, could be identified at Rank-1 based on his probe images released by the authorities, the older brother could only be identified at Rank-12,446 (from a gallery of one million images with no demographic filtering). Figure 16 shows three gallery face images of Tamerlan Tsarnaev (*1x*, *1y*, and *1z* [34]) and two probe face images (*1a* and *1b*) which were released by the FBI during the investigation.[9] Because the probe images of Tamerlan Tsarnaev are of poor quality, particularly due to wearing of sunglasses and a hat, we also asked a sketch artist to draw a sketch of Tamerlan Tsarnaev (*1c* in Fig. 16) while viewing the two probe images.[10]

To simulate a large-scale forensic investigation, the three gallery images of Tamerlan Tsarnaev were added to a background set of one million mugshot images of 324,696 unique subjects from the PCSO database. Particularly due to the occlusion of eyes, the probe images are difficult for the COTS face matcher to identify (though they can be enrolled with manually marked eye locations), as shown in Table VI. However, the retrieval rank for the sketch (*1c* in Fig. 16) is much better compared to the two probe images (*1a* and *1b* in Fig. 16), with the best match at Rank-6,259 for max fusion of multiple images of Tamerlan Tsarnaev (*1x*, *1y*, and *1z*) in the gallery. With demographic filtering [35] (white male in the age range of 20 to 30 filters the gallery to 54, 638 images of 13, 884 subjects), the sketch is identified with gallery image *1x* (a mugshot)[11] in Fig. 16 at Rank-112. Again, score fusion of multiple images per subject in the gallery further lowers the retrieval to Rank-71. The entire media collection (here, *1a*, *1b*, and *1c* in Fig. 16) is matched at Rank-82 against the demographic-filtered and multiple image-fused gallery.

E. Watch List Scenario: Open Set Identification

We report the DIR vs. FAR curves of open set identification in Figs. 17(a) and (b). With a single image or single video per subject in the gallery, the DIR values at 1% FAR are about 25% and 10% for still image probe and video clip probe, respectively. This suggests that a large percentage of probe images or video clips that are matched to their gallery true mates at a low rank in a closed set identification scenario, can no longer be successfully matched in an open set scenario.

[9] http://www.fbi.gov/news/updates-on-investigation-into-multiple-explosions-in-boston

[10] *"I was living in Costa Rica at the time that event took place and while I saw some news coverage, I didn't see much and I don't know what he actually looks like. The composite I am working on is* 100% *derived from what I am able to see and draw from the images you sent. I can't make up information that I can't see, so I left his hat on and I can only hint at eye placement."* - Jane Wankmiller, forensic sketch artist, Michigan State Police.

[11] http://usnews.nbcnews.com/_news/2013/05/06/18086503-funeral-director-in-boston-bombing-case-used-to-serving-the-unwanted?lite

TABLE V

CLOSED SET IDENTIFICATION ACCURACIES (%) FOR MATCHING CONSOLIDATED 3D FACE MODELS BUILT FROM (*a*) ALL FRAMES OF A VIDEO TRACK OR (*b*) A SUBSET OF HIGH QUALITY (HQ) VIDEO FRAMES

Consolidated 3D Model: All Frames

	R-1	R-20	R-100	R-200
s_3	33.1	54.1	67.3	72.8
s_4	29.4	51.7	64.8	71.1
sum	34.6	56.4	68.2	74.1

(a)

Consolidated 3D Model: Frame Selection

	R-1	R-20	R-100	R-200
s_3	34.4	56.6	67.8	73.4
s_4	29.8	52.4	66.5	72.7
sum	35.9	58.3	69.9	75.1

(b)

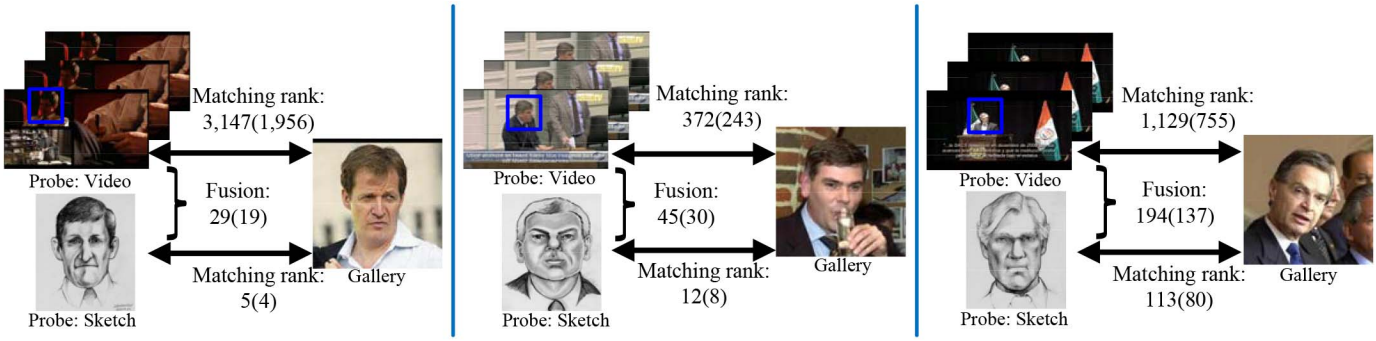


Fig. 15. Three examples where the face sketches drawn by a forensic artist after viewing the low-quality videos improve the retrieval rank. The retrieval ranks without and with combining the demographic information (gender and race) are given in the form of #(#).

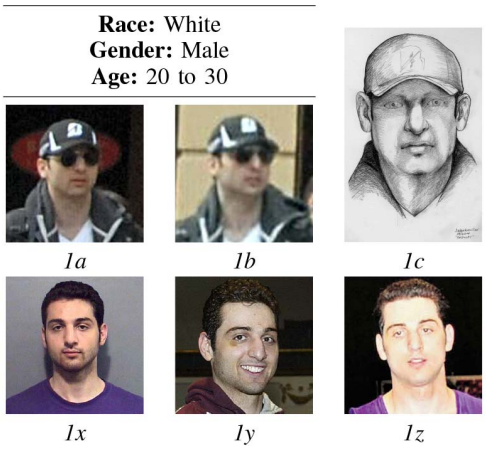


Fig. 16. Face images used in our case study on identification of Tamerlan Tsarnaev, one of the two suspects of the 2013 Boston Marathon bombings. Probe (*1a*, *1b*) and gallery (1x, 1y, and 1z) face images are shown. *1c* is a face sketch drawn by a forensic sketch artist after viewing *1a* and *1b*, and a low quality video frame from a surveillance video.

Of course, this comes at the benefit of much lower false alarms than in the closed set identification. The proposed face media collection based matching still shows improvement over single media based matching. For example, at 1% FAR, face media collection based matching leads to about 20% and 15% higher DIRs for still image and video clip probes, respectively.

F. Large Gallery Results

In order to simulate the large-scale nature of operational face identification, we extend the size of our gallery by including *one million* face images from the PCSO database. We acknowledge that there may be a bias towards matching between LFW probe and LFW gallery images versus matching LFW probe with PCSO gallery images. This bias is likely due to the fact that the gallery face images in LFW are not necessarily frontal with controlled illumination, expression, etc., while the background face images from PCSO are mugshots of generally cooperative subjects.

The extended gallery set with 1M face images makes the face identification problem more challenging. Figure 17(c) gives the media collection based face identification accuracies

TABLE VI

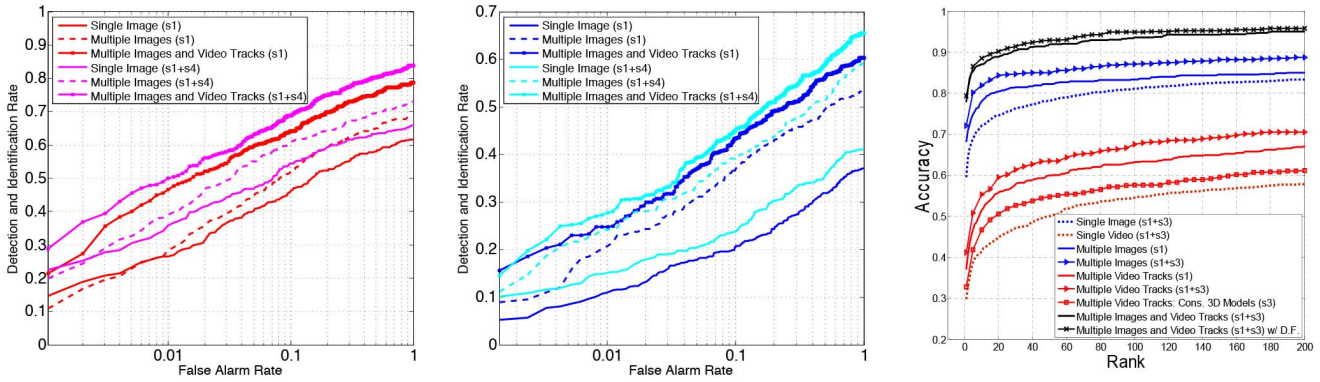
RETRIEVAL RANKS FOR PROBE IMAGES (*1a, 1b*) AND SKETCH (*1c*) MATCHED AGAINST GALLERY IMAGES *1x, 1y,* AND *1z* WITH AN EXTENDED SET OF ONE MILLION MUG SHOTS (*a*) WITHOUT AND (*b*) WITH DEMOGRAPHIC FILTERING. ROWS MAX AND MEAN DENOTE SCORE FUSION OF MULTIPLE IMAGES OF THIS SUSPECT IN THE GALLERY; COLUMNS MAX AND SUM ARE SCORE FUSION OF THE THREE PROBES

(*a*)

	1a	*1b*	*1c*	max	sum
1x	117,322	475,769	**8,285**	18,710	27,673
1y	12,444	440,870	63,313	38,298	28,169
1z	87,803	237,704	53,771	143,389	55,712
max	9,409	117,623	**6,259**	14,977	**6,281**
mean	13,658	125,117	8,019	20,614	8,986

(*b*)

	1a	*1b*	*1c*	max	sum
1x	5,432	27,617	**112**	114	353
1y	518	25,780	1,409	1,656	686
1z	3,958	14,670	1,142	2,627	1,416
max	374	6,153	94	109	106
mean	424	5,790	**71**	109	**82**



(a) Probe: Single Image (b) Probe: Single Video Track (c) Large Gallery

Fig. 17. Scenarios of open set and closed set identifications. Open set identification with (a) a single face image as the probe and various media collections as the gallery and (b) a single face video track as the probe and various media collections as the gallery; the legend denotes the gallery media collection in (a) and (b). Closed set identification of (c) various media collections as probe against a large gallery set with one million background face images from the PCSO database; the legend denotes the probe media collection; the black curve denoted with "D.F." indicates that demographic information (gender and race) is also fused with the other face media. Note that the ordinate scales are different in (a) and (b) to accentuate the difference among the plots.

with $1M$ background face images. A comparison between Fig. 17(c) and Fig. 9 shows that the proposed face media collection based matching generalizes well to a large gallery set.

VII. Conclusions

In this paper, we studied face identification of persons of interest in unconstrained imaging scenarios with uncooperative subjects. Given a face media collection of a person of interest (*i.e.,* face images and video clips, 3D face models built from image(s) or video frame(s), face sketch, and demographic information), we have demonstrated an incremental improvement in the identification accuracy of a COTS face matching system. We believe this is of great value to forensic investigations and "lights out" watch list operations, as matching the entire probe collection outputs a *single* ranked list of candidate identities, rather than a ranked list for each face media sample. Evaluations are provided in the scenarios of closed set identification, open set identification, closed set identification with a large gallery, and verification. Our contributions can be summarized as follows:

1) A collection of face media, such as image, video, 3D face model, face sketch, and demographic information, on a person of interest improves identification accuracies, on average, particularly when individual face media samples are of low quality for face matching.
2) Pose correction of unconstrained 2D face images and video frames (via 3D face modeling) prior to matching improves the accuracy of a state of the art COTS face matcher. This improvement is especially significant when match scores from rendered pose corrected images are fused with match scores from original face imagery.
3) While a single consolidated 3D face model can summarize the entire video track, matching all the pose corrected frames of a video track performs better than the consolidated model.
4) Quality based fusion of match scores of different media types performs better than fusion without incorporating the quality.
5) The value of forensic sketch drawn based on low quality videos or low quality images of the suspect is demonstrated in the context of one of the Boston bombing suspects and YTF video tracks.

Pose-corrected face images from the LFW database, pose-corrected video frames from the YTF database, forensic sketches, and experimental protocols used in this paper have been made publicly available. Our ongoing work involves investigation of more effective face quality measures to further boost the performance of fusion for matching a media collection. A reliable face quality value will prevent forensic analysts from having to attempt all possible combinations of face media matching. Another important problem is to improve 3D model construction from multiple still images or video frames.

Acknowledgements

The face sketches were prepared by Jane Wankmiller and Sarah Krebs, sketch artists with the Michigan State Police.

References

[1] A. K. Jain, B. Klare, and U. Park, "Face matching and retrieval in forensics applications," *IEEE Multimedia*, vol. 19, no. 1, p. 20, Jan. 2012.
[2] (Nov. 2013). *IARPA Broad Agency Announcement: BAA-13-07, Janus Program*. [Online]. Available: http://www.iarpa.gov/Programs/sc/Janus/solicitation_janus.html
[3] S. Z. Li and A. K. Jain, Eds., *Handbook of Face Recognition*, 2nd ed. New York, NY, USA: Springer-Verlag, 2011
[4] 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," Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.
[5] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 529–534.
[6] H. Wang, B. Kang, and D. Kim, "PFW: A face database in the wild for studying face identification and verification in uncontrolled environment," in *Proc. 2nd IAPR Asian ACPR*, Nov. 2013, pp. 356–360.
[7] E. G. Ortiz and B. C. Becker, "Face recognition for web-scale datasets," *Comput. Vis. Image Understand.*, vol. 118, pp. 153–170, Jan. 2014.
[8] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[9] National Institute of Standards and Technology (NIST). (Jun. 2013). *Face Homepage*. [Online]. Available: http://face.nist.gov
[10] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *Proc. BMVC*, 2013, pp. 1–12.
[11] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 3025–3032.
[12] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 1701–1708.
[13] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 1891–1898.
[14] S. Liao, Z. Lei, D. Yi, and S. Z. Li, "A benchmark study of large-scale unconstrained face recognition," in *Proc. IAPR/IEEE IJCB*, Sep./Oct. 2014, pp. 1–8.
[15] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–8.
[16] L. Best-Rowden, B. Klare, J. Klontz, and A. K. Jain, "Video-to-video face matching: Establishing a baseline for unconstrained face recognition," in *Proc. IEEE 6th Int. Conf. BTAS*, Sep./Oct. 2013, pp. 1–8.
[17] G. Sharma, S. ul Hussain, and F. Jurie, "Local higher-order statistics (LHS) for texture categorization and facial analysis," in *Proc. 12th ECCV*, 2012, pp. 1–12.
[18] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 3499–3506.
[19] 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 *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 3554–3561.
[20] S. Liao, A. K. Jain, and S. Z. Li, "Partial face recognition: Alignment-free approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1193–1205, May 2013.
[21] E. Mostafa, A. Ali, N. Alajlan, and A. Farag, "Pose invariant approach for face recognition at distance," in *Proc. 12th ECCV*, 2012, pp. 15–28.
[22] X. Ge, J. Yang, Z. Zheng, and F. Li, "Multi-view based face chin contour extraction," *Eng. Appl. Artif. Intell.*, vol. 19, no. 5, pp. 545–555, Aug. 2006.
[23] Y. Lin, G. Medioni, and J. Choi, "Accurate 3D face reconstruction from weakly calibrated wide baseline images with profile contours," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 1490–1497.
[24] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith, "Fully automatic pose-invariant face recognition via 3D pose normalization," in *Proc. IEEE ICCV*, Nov. 2011, pp. 937–944.
[25] C. P. Huynh, A. Robles-Kelly, and E. R. Hancock, "Shape and refractive index from single-view spectro-polarimetric images," *Int. J. Comput. Vis.*, vol. 101, no. 1, pp. 64–94, Jan. 2013.
[26] H. Han, C. Otto, and A. K. Jain, "Age estimation from face images: Human vs. machine performance," in *Proc. ICB*, Jun. 2013, pp. 1–8.
[27] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE 12th ICCV*, Oct. 2009, pp. 365–372.
[28] K. T. Taylor, *Forensic Art and Illustration*. Boca Raton, FL, USA: CRC Press, 2000.
[29] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain, "Matching composite sketches to face photos: A component-based approach," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 1, pp. 191–204, Jan. 2013.
[30] S. Klum, H. Han, A. K. Jain, and B. Klare, "Sketch based face recognition: Forensic vs. composite sketches," in *Proc. ICB*, Jun. 2013, pp. 1–8.
[31] A. A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics*. New York, NY, USA: Springer-Verlag, 2006.
[32] U. Uludag and A. K. Jain, "Attacks on biometric systems: A case study in fingerprints," *Proc. SPIE*, vol. 5306, pp. 622–633, Jun. 2004.
[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[34] J. C. Klontz and A. K. Jain, "A case study on unconstrained facial recognition using the Boston marathon bombings suspects," Michigan State Univ., Lansing, MI, USA, Tech. Rep. MSU-CSE-13-4, May 2013.
[35] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 6, pp. 1789–1801, Dec. 2012.

Lacey Best-Rowden (S'14) received the B.S. degree in computer science and mathematics from Alma College, Alma, MI, USA, in 2010. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA. Her research interests include pattern recognition, computer vision, and image processing with applications in biometrics.

Hu Han (M'13) is currently a Research Associate with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA. He received the B.S. degree in computer science from Shandong University, Jinan, China, in 2005, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2011. His research interests include computer vision, pattern recognition, and image processing, with applications to biometrics, forensics, law enforcement, and security systems.

Charles Otto (S'13) received the B.S. degree from the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA, in 2008. He was a Research Engineer at IBM, Armonk, NY, USA, from 2006 to 2011. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Michigan State University. His research interests include pattern recognition, image processing, and computer vision, with applications to face recognition.

Brendan F. Klare (M'11) received the B.S. and M.S. degrees from the University of South Florida, Tampa, FL, USA, in 2007 and 2008, respectively, and the Ph.D. degree from Michigan State University, East Lansing, MI, USA, in 2012, all in computer science. He is currently a Lead Scientist with Noblis, Falls Church, VA, USA. From 2001 to 2005, he served as an Airborne Ranger Infantryman in the 75th Ranger Regiment. His research interests include pattern recognition, image processing, and computer vision. He has authored several papers on the topic of face recognition. He was a recipient of the Honeywell Best Student Paper Award at the 2010 IEEE Conference on Biometrics: Theory, Applications, and Systems.

Anil K. Jain (LF'14) is currently a University Distinguished Professor with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA. His research interests include pattern recognition and biometric authentication. He served as the Editor-in-Chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (1991-1994). He has coauthored a number of books, including *Handbook of Fingerprint Recognition* (2009), *Handbook of Biometrics* (2007), *Handbook of Multibiometrics* (2006), *Handbook of Face Recognition* (2011), *Biometrics: Personal Identification in Networked Society* (1999), and *Algorithms for Clustering Data* (1988). He served as a member of the Defense Science Board and the National Academies Committees on Whither Biometrics and Improvised Explosive Devices. He was a recipient of the 1996 IEEE TRANSACTIONS ON NEURAL NETWORKS OUTSTANDING PAPER AWARD and the Pattern Recognition Society Best Paper Awards in 1987, 1991, and 2005. He received the Fulbright Award, the Guggenheim Award, the Alexander von Humboldt Award, the IEEE Computer Society Technical Achievement Award, the IEEE Wallace McDowell Award, the ICDM Research Contributions Award, and the International Association for Pattern Recognition (IAPR) King-Sun Fu Award. He is a fellow of the American Association for the Advancement of Science, the Association for Computing Machinery, IAPR, and the Society for Optics and Photonics.