Learning Face Image Quality from Human Assessments

Lacey Best-Rowden, *Member, IEEE,* and Anil K. Jain, *Life Fellow, IEEE*

***Abstract*—Face image quality can be defined as a measure of the utility of a face image to automatic face recognition. In this work, we propose (and compare) two methods for learning face image quality based on target face quality values from (i) human assessments of face image quality (matcher-independent), and (ii) quality values computed from similarity scores (matcher-dependent). A support vector regression model trained on face features extracted using a deep convolutional neural network (ConvNet) is used to predict the quality of a face image. The proposed methods are evaluated on two unconstrained face image databases, LFW and IJB-A, which both contain facial variations encompassing a multitude of quality factors. Evaluation of the proposed automatic face image quality measures shows we are able to reduce the FNMR at 1% FMR by at least 13% for two face matchers (a COTS matcher and a ConvNet matcher) by using the proposed face quality to select subsets of face images and video frames for matching templates (i.e., multiple faces per subject) in the IJB-A protocol. To our knowledge, this is the first work to utilize human assessments of face image quality in designing a predictor of unconstrained face quality that is shown to be effective in cross-database evaluation.**

***Index Terms*—face image quality, face recognition, biometric quality, crowdsourcing, unconstrained face images.**

I. Introduction

THE performance of automatic face recognition systems largely depends on the quality of the face images acquired for comparison. Under controlled image acquisition conditions (*e.g.,* ID card face images) with uniform lighting, frontal pose, neutral expression, and standard image resolution, face recognition systems can achieve extremely high accuracies. For example, the NIST MBE [1] reported face verification accuracy of >99% True Accept Rate (TAR) at 0.1% False Accept Rate (FAR) for a database of visa photos, and the NIST FRVT 2013 [2] reported 96% rank-1 identification accuracy for a database of law enforcement face images (*e.g.,* mugshots). However, many emerging applications of face recognition seek to operate on face images captured in less than ideal conditions (*e.g.,* surveillance) where large intra-subject facial variations are more prevalent, or even the norm, and can significantly degrade recognition accuracy. The NIST FRVT Ongoing[1] currently shows FNMRs of less than 3% at 0.01% FMR for the top performing algorithms, while at 0.1% FMR, an order of magnitude higher, the FNMRs of the same algorithms jumps to 23% or greater when evaluated on "in the wild" face images.

Fig. 1. Cropped faces from frames of a sample video in the IJB-A [6] unconstrained face database; faces are sorted from high to low face quality by the proposed MQV predictor.

The performance of biometric recognition, in general, is driven by the quality of biometric samples (*e.g.,* fingerprint, iris, and face). Biometric sample quality is defined as *a measure of a sample's utility to automatic matching* [3]–[5]. A biometric quality measurement should be an indicator of recognition performance where correlation with error rates, such as false non-match rate (FNMR), false match rate (FMR), or identification miss rate, is a desirable property. Biometric samples determined to be of poor quality should cause a recognition system to fail. In this work, we are only focusing on face quality. Biometric quality for other modalities (*e.g.,* fingerprint and iris) are summarized and reviewed in [4], [5].

Automatic prediction of *face* quality (prior to matching and recognition) can be useful for a number of practical applications. A system with the ability to detect poor quality face images can subsequently process them accordingly. In negative identifications systems (*e.g.,* automated security checkpoints at airports), persons may intentionally present low quality face images to the system to evade recognition; face quality assessment could flag such attempts and deny services (*e.g.,* entry through the checkpoint). Face image quality can also be used for quality-based fusion when multiple face images (*e.g.,* sequence of video frames, see Fig. 1) and/or biometric modalities [7], [8] (*e.g.,* face and fingerprint [9]) of the same subject are available, as well as for 3D face modeling from a collection of face images [10]. Additionally, dynamic recognition approaches [11] can make use of face image quality where high-quality face images can be assigned

L. Best-Rowden was with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824 USA. She is now with the Risk and Authentication Products Organization at Visa Inc., Foster City, CA 94404 USA. (email: lbestrow@visa.com)

A. K. Jain is with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, 48824 USA. (email: jain@cse.msu.edu)

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org/Xplore/home.jsp, provided by the author. The material includes the human assessments of face quality collected for this study via crowdsourcing on Amazon Mechanical Turk. Contact lbestrow@visa.com for further questions about this work.

[1]https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing [accessed October 3, 2017]

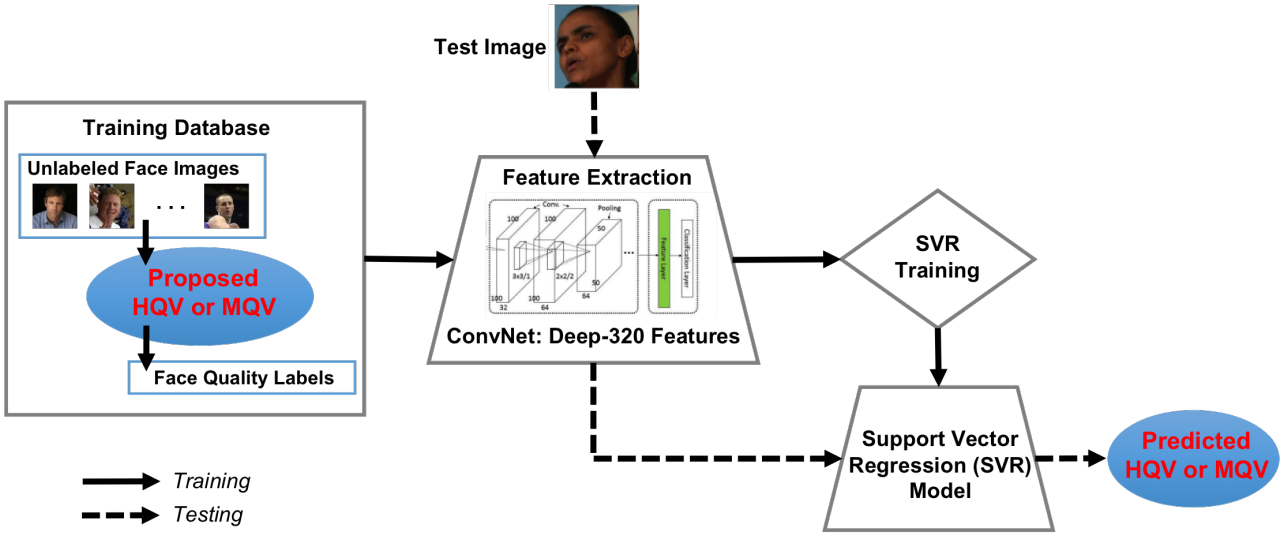


Fig. 2. Diagram of the proposed framework (training and testing) for face quality prediction using a support vector regression model (SVR) trained on ConvNet face features extracted from a training database for face quality. The key contributions of this work are two different methods for labeling face quality to construct a training database based on (i) Human Quality Values (HQV) and (ii) Matcher Quality Values (MQV). The HQV method labels the training database using human quality assessments of face image pairs crowdsourced via Amazon Mechanical Turk, whereas the MQV method labels the training database using only similarity scores from face matchers. Hence, in this work, we propose and compare matcher-independent (*i.e.,* HQV) and matcher-dependent (*i.e.,* MQV) methods for face quality prediction. The predicted quality values are evaluated and compared in terms of their utility to improving face recognition performance of either COTS face matchers or a face matcher based on the same ConvNet features used for face quality prediction. (The COTS matchers are "black-box" but they are also based on convolutional neural networks.)

to high-throughput algorithms, while low-quality face images are assigned to slower, but more robust, algorithms.

Because biometric sample quality is defined in terms of automatic recognition performance, human visual perception of image quality may not be well correlated with recognition performance [3]–[5]. Particularly, given a fingerprint or iris image, it is difficult for a human to assess the quality in the context of recognition because humans (excluding forensic experts) do not naturally use fingerprints or iris textures for person recognition. However, the human visual system is extremely advanced when it comes to recognizing the faces of individuals, a routine daily task. In fact, it was recently shown that humans surpass the performance of current state-of-the-art automated systems on recognition of challenging, low-quality, face images [12]–[14]. Even so, to the best of our knowledge, only a few studies have investigated face quality assessment by humans. Adler and Dembinsky [15] found weak correlation between human and algorithm measurements of face quality (98 mugshots of 29 subjects, eight human evaluators), while Hsu *et al.* [16] found some consistency between human perception and recognition-based measures of face quality (frontal and controlled illumination face images, two human evaluators).

Recent works on automatic face recognition have devoted efforts towards recognition of *unconstrained* facial imagery [6], [17]–[21] where facial variations of any kind can be simultaneously present (*e.g.,* face images from surveillance cameras [13], [22]). However, prior work in face image quality has primarily focused on the quality of lab-collected face image databases (*e.g.,* FRGC [23], GBU [24], Multi-PIE [25]) where facial variations such as illumination and pose are synthetic/staged/simulated in order to isolate and facilitate evaluation of the different quality factors.

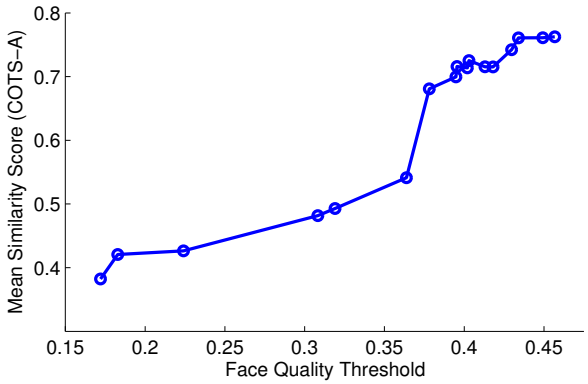
In this work, we focus on automatic face image quality of unconstrained face images using the Labeled Faces in the Wild (LFW) [26] and IARPA Janus Benchmark A (IJB-A) [6] unconstrained face datasets. The contributions of this work are summarized as follows:

- Collection of human ratings of face image quality for a large database of unconstrained face images (namely, LFW [26]) by crowdsourcing a small set of pairwise comparisons of face images and inferring the complete ratings with matrix completion [27].
- Investigation of the utility of face quality assessment by humans in the context of automatic face recognition performance. This is the first study on human quality assessment of face images that exhibit a wide range of quality factors (*i.e., unconstrained* face images).
- A model for automatic prediction of face image quality trained on convolutional neural network (ConvNet) face features extracted from a training database labeled for the face quality task. See Fig 2.
- We propose and compare two different methods for establishing the labels of a training database for the face quality task. The two methods are based on: (i) human quality ratings (*matcher-independent*) and (ii) quality values computed from similarity scores obtained from face matchers (*matcher-dependent*). See Fig. 2.

Our experimental evaluation follows the methodology advocated by Grother and Tabassi [3] where a biometric quality measurement is tested by "relating quality values to empirical matching results." The quantitative evaluation presented is aimed at the application of using face image quality to improve error rates (*e.g.,* FNMR) of automatic face recognition

(a)

(b)



(c)

Fig. 3. (a) A gallery and (b) a probe template of the same subject from the IJB-A database [6]. Face image quality values automatically predicted by the proposed HQV predictor are given in green (lower value indicates lower face quality). To obtain a single similarity score for the multiple faces in the gallery and probe templates, score-level fusion is typically the baseline approach. (c) Score-level fusion (mean rule) of COTS-A similarity scores using only those faces with face quality above a threshold increases the fused similarity score as the threshold increases. In this scenario, a monotonically increasing relationship is desired between the mean genuine similarity score and the face quality threshold because higher quality faces should result in higher genuine similarity.

systems by rejecting low-quality face images. For example, in template-based matching (*e.g.,* the IJB-A protocol [6]) standard score-level fusion over multiple faces per subject can be improved by removing low-quality faces prior to computing the mean of the similarity scores (see Fig. 3).

II. Related Work

Countless studies (*e.g.,* [31], [36], [37]) have analyzed the performance of face recognition algorithms with respect to different covariates such as pose, illumination, expression, resolution, and others. In doing so, knowledge about face quality as it pertains to recognition performance has helped to guide innovative solutions over the years. Given the known sensitivities of recognition performance when faces deviate from constrained conditions, earlier works proposed the quality of a face as its similarity to reference, or "ideal", face images (typically frontal pose, uniform illumination, neutral expression). For example, [38] uses luminance distortion from a high quality reference image for adaptive fusion of two face representations, Best-Rowden *et al.* [39] investigated structural similarity (SSIM) for quality-based fusion within a collection of face media, and Wong *et al.* [40] propose probabilistic similarity to a reference model of ideal face images for selecting high quality frames in video face verification. While reference-based face quality is easily interpretable, these approaches depend on the face images chosen as reference and may not generalize to truly unconstrained faces from different databases.

Video-based face recognition, *e.g.,* surveillance scenarios [13], [22], is one of the primary applications of face quality. Using face quality for frame selection not only reduces storage and computation time required to process the volume of imagery in a video, but can also improve recognition performance. For example, Goswami *et al.* [41] proposed highest visual entropy for frame selection from videos, and more recently, Goswami *et al.* also proposed "feature-richness" based on entropy in the wavelet domain [42] to achieve high verification accuracy on the PaSC video face dataset [13], [22].

Also worth mentioning are recent works on video-based, or more generally, *template*-based (multiple images and/or videos per subject) face recognition which train a network to learn a single (fixed-length) face representation from a variable number of video frames (and/or images) of a given subject [21], [43]. Although these methods are trained specifically for the purpose of face verification or identification, a byproduct of the training process is a measure of face quality because the weights or coefficients learned for combining the multiple faces into a single representation tend to reflect the quality of a face for recognition purposes. For example, Tran *et al.* [21] propose a Disentangled Representation learning-Generative Adversarial Network (DR-GAN), which, when trained to learn a single representation from multiple images of a subject, gives confidence coefficients that indicate the face quality. Similarly, Yang *et al.* [43] propose a Neural Aggregation Network (NAN) for video face recognition that consists of a CNN that computes features for individual faces and attention blocks that learn weights for pooling the features from a subject's multiple video frames and images into a single representation. Both Tran *et al.* [21] and Yang *et al.* [43] show that the proposed architectures improve face recognition performance over baseline fusion schemes and give visual examples that suggest face quality is a promising byproduct of training for face recognition tasks.

Most similar to our work are learning-based approaches where the target face quality is first defined in some manner to be indicative of recognition performance. The target quality value can be a prediction of the genuine score [16], [28], [34], a bin indicating that an image is poor/fair/good for matching

TABLE I
SUMMARY OF RELATED WORK ON AUTOMATIC METHODS FOR FACE IMAGE QUALITY

Study (year)	Database: Num. of images (subjects)	Target Quality Value	Learning Approach	Evaluation
Hsu *et al.* [16] (2006)	FRGC: 1,886 (n/a) passports: 2,000 (n/a) mugshots: 1,996 (n/a)	Continuous (genuine score)	Neural network to combine 27 quality measures (exposure, focus, pose, illumination, etc.) for prediction of genuine scores	ROC curves for different levels of quality (FaceIt algorithm by Identix)
Aggarwal *et al.* [28] (2011)	Multi-PIE: 6,740 (337)* FacePix: 1,830 (30)	Continuous (genuine score) or Binary (algorithm success vs. failure; requires matching prior to quality)	MDS to learn a mapping from illumination features to genuine scores. Predicted genuine score compared to algorithm score to decide algorithm success or failure	Prediction accuracy of algorithm success vs. failure, ROC curves for predicted, actual, 95% and 99% retained (SIFT-based and PittPatt algorithms)
Phillips *et al.* [29] (2013)	PaSC: 4,688 (n/a) GU†: 4,340 (437)	Binary (low vs. high)	PCA + LDA classifier	Error vs. Reject curve for FNMR vs. percent of images removed
Bharadwaj *et al.* [30] (2013)	CAS-PEAL: n/a (1,040) SCface: n/a (130)	Quality bins (poor, fair, good, excellent)	Multi-class SVM trained to predict face quality bin from holistic face features (GIST and HOG)	ROC curves, rank-1 accuracy, EER, % histogram overlap (COTS algorithm)
Abaza *et al.* [31] (2014)	GU†: 4,340 (437)	Binary (good vs. ugly)	Neural network (1-layer) to combine contrast, brightness, sharpness, focus, and illumination measures	Rank-1 identification for blind vs. quality-selective fusion
Dutta *et al.* [32] (2014)	Multi-PIE: 3,370 (337)‡	Continuous (false reject rate)	Probability density functions (PDFs) model interaction between image quality (deviations from frontal and uniform lighting) and recognition performance	Predicted vs. actual verification performance for different clusters of quality (FaceVACS algorithm)
Kim *et al.* [33] (2015)	FRGC: 10,448 (322)	Binary (low vs. high) or Continuous (confidence of the binary classifier)	Objective (pose, blurriness, brightness) and Relative (color mismatch between train and test images) face image quality measures as features fed into AdaBoost binary classifier	Identification rate w.r.t. fraction of images removed, ROC curve with and without low quality images (SRC face recognition algorithm)
Vignesh *et al.* [34] (2015)	ChokePoint: 48 videos (25)	Continuous (genuine score)	CNN is used to predict the genuine score from a matcher based on LBP/HOG features and Mutual Subspace Method (MSM) for image set matching	Face subset selection for input to MSM to improve video-based face verification performance
Chen *et al.* [35] (2015)	SCface: 2,080 (130) (trained with FERET, FRGC, LFW, and non-face images)	0 – 100 (rank-based quality score)	A ranking function is learned by assuming images from different databases are of different quality and images from the same database are of equal quality	Visual quality-based rankings, Identification rate (Gabor filter based matcher)
Proposed Approach	LFW: 13,233 (5,749) for training and testing IJB-A: 5,712 images and 2,085 videos (500) for testing	Continuous (human quality ratings or normalized comparison scores)	Support vector regression with image features from a deep convolutional neural network [17]	Error vs. Reject curves, quality-based subset selection for template matching, visual quality-based ranking

Note: n/a *indicates that the authors did not report the number of images or subjects (an unknown subset of the database may have been used).*
*Only the illumination subset of Multi-PIE database [25] was used. †GU denotes the Good and Ugly partitions of the Good, Bad, and Ugly (GBU) face database [24]. ‡Only neutral expression face images from Multi-PIE database [25] were used.

[30], or a binary value of low vs. high quality image [29], [31], [33]. For example, Bharadwaj *et al.* fuse similarity scores from two COTS matchers, define quality bins based on CDFs of faces that were matched correctly and incorrectly, and use a support vector machine (SVM) trained on holistic image features to classify the quality bin of a test image [30]. Vignesh *et al.* [34] use a four-layer CNN to predict the genuine scores from Mutual Subspace Method (MSM) for image set matching with LBP/HOG features. Table I summarizes these approaches.

While most related methods in Table I define the target quality values for a training database of face images, Chen *et al.* [35] instead propose using Parikh and Grauman's learning to rank framework [44] for rank-based face quality. Chen *et al.*'s framework assumes (i) a rank-ordering ($\prec$) of a set of databases, such that (non-face images) $\prec$ (unconstrained face images) $\prec$ (ID card face images), and (ii) face images from the same database have equal quality; rank weights are learned using five different image features and then mapped to a quality score $0-100$ [35].

In this work, we establish the target face quality values (defined as either human quality ratings or score-based values from a face matcher) of a large database of unconstrained face images, extract image features using a deep ConvNet [17], and learn a model for prediction of face quality from the

ConvNet features using support vector regression (SVR). Our approach is most similar to [30], but our target quality values are continuous, allowing for a fine-tuned quality-based ranking of a collection of face images. Additionally, [30], nor any other methods in Table I, does not investigate target quality defined from human quality assessments as we do in this paper.

III. Face Databases and Matchers

This work utilizes three unconstrained face databases: CASIA-WebFace [19], Labeled Faces in the Wild (LFW) [26], and IARPA Janus Benchmark-A (IJB-A) [6]. The CASIA database consists of 494,414 images of 10,575 subjects, LFW consists of 13,233 images of 5,749 subjects, and IJB-A consists of 5,712 images and 2,085 videos of 500 subjects. All three databases were compiled by crawling the internet for faces of celebrities and public figures. Faces in the LFW database were detected by the Viola-Jones face detector [45], so the pose variations are limited by the pose tolerance of the Viola-Jones detector. Faces in IJB-A were manually located, so the database is considered more challenging than LFW due to full pose variations [6]. Fig. 4 shows sample face images from these two databases. The CASIA database has been commonly used to train deep neural networks for face recognition [17], [19]–[21], [46]–[49]. The two proposed predictors of face quality, MQV and HQV, are both trained using CASIA and LFW and evaluated on the IJB-A database (also evaluated on LFW using cross-fold validation).

This work also utilizes three different face matchers: two commercial face matchers, denoted as COTS-A and COTS-B[2,3], and a deep convolutional neural network (ConvNet). The ConvNet matcher is based on the network architecture in [17] and is trained on the CASIA database. See Table II for an initial performance comparison of the three matchers on the BLUFR protocol [50] for the LFW database. Whereas [17] uses multiple ConvNet models trained on different facial subregions (eyes, nose, mouth, etc.) to boost performance for face recognition, we instead use the feature representation from a single ConvNet in this work. Although the three face matchers used in this work (COTS-A, COTS-B, and ConvNet) are all based on deep convolutional networks (CNNs)[4], we use multiple matchers to demonstrate the applicability of the proposed face quality methods to more than just a single matcher. The MQV

Each face matcher is used to (i) establish target face quality values using the matcher's comparison scores for training the proposed MQV predictor and (ii) evaluate the utility of both the MQV and HQV predictors for face recognition purposes. In addition to using the ConvNet matcher for the two aforementioned purposes, the feature representation from the ConvNet is also used as the feature representation for the proposed face quality predictors. More details about the ConvNet architecture are given in Sec. V.

[2]COTS-A was one of the top performers in the 2013 NIST FRVT [2] and COTS-B is currently competitive with the top algorithms in the NIST FRVT Ongoing.

[3]The COTS-A SDK does not include any face quality measures, but the COTS-B SDK includes a face quality measure that indicates the expected FNMR of an image. We compare the proposed face quality with COTS-B face quality in Section VI-B1.

[4]COTS-A and COTS-B are "black-box" to us but the COTS vendors have disclosed that they are based on deep CNNs.

(a) (b)

Fig. 4. Sample face images from the (a) LFW [26] and (b) IJB-A [6] unconstrained face databases.

TABLE II
Verification and Open-Set Identification Performance for the BLUFR Protocol [50] of the LFW Database [26]

Algorithm	TAR @ 0.1% FAR	DIR† @ 1% FAR
HDLBP + JointBayes [18]*	41.7	18.1
Yi *et al.* [19]	80.3	28.9
ConvNet [17] (# nets = 1)	85.0	49.1
ConvNet [17] (# nets = 9)	89.8	55.9
COTS-A	88.1	76.3
COTS-B	76.0	53.2

*Performance here for [18] was reported by [50]

†DIR = Detection and Identification Rate

IV. Face Image Quality Labels

Biometrics and computer vision heavily rely on supervised learning techniques when training sets of *labeled* data are available. With the aim of developing an automatic method for face quality, compiling a quality-labeled face image database is not straightforward. The definition of face image quality (*i.e.,* a predictor of automatic matching performance) does not lend itself to explicit labels of face image quality, unlike labels of facial identity or face vs. non-face labels for face recognition and detection tasks, respectively. Possible approaches for establishing target quality labels of face images include:

(i) Combine various measurements of image quality factors into a single value for overall face quality.
(ii) Human annotations of perceived image quality.
(iii) Use comparison scores (or performance measures) from automatic face recognition matchers.

The issue with (i) is that it is an *ad-hoc* approach and, thus far, has not achieved much success (see [29]). The issue with (ii) is that human perception of quality may not be indicative of automatic recognition performance; previous works [3], [30] have stated this conjecture, but to our knowledge, the only studies to investigate human perception of face quality were conducted on constrained face images (*e.g.,* mugshots) [15], [16]. The issue with (iii) is that comparison scores are obtained from a *pair* of images, so labeling single images based on comparison scores (or performance) can be problematic. However,

this approach achieved some success for fingerprint quality [3], [51], and only few studies [29], [30] have considered it for face quality. In this work, we investigate both methods 2) and 3), detailed in the remainder of this section.

A. Human Quality Values (HQV)

Because of the inherent ambiguity in the definition of face image quality, framing an appropriate prompt to request a human to label the quality of a face image is challenging. If asked to rate a single face image on a scale of 1 to 5, for example, there are no notions as to the meaning of the different quality levels. Prior exposure to the variability in the face images that the human will encounter may be necessary so that they know what kinds of "quality" to expect in face images (*i.e.,* a baseline) before beginning the quality rating task. Crowdsourcing literature [27] has demonstrated that ordinal (comparison-based) tasks are generally easier for participants and take less time than cardinal (score-based) tasks. Ordinal tasks additionally avoid calibration efforts needed for cardinal responses from raters inherently using different ranges for decision making (*i.e.,* biased ratings, inflated vs. conservative ratings, changes in absolute ratings from exposure to more data). For these reasons, we choose to collect relative pairwise comparisons of face image quality by asking the following question: "Which face (left or right) has better quality?"

Given the collected pairwise face comparisons, to obtain absolute quality ratings for *individual* face images, we make use of a matrix completion approach [27] to infer the quality rating matrix from the pairwise comparisons. Because it is infeasible to have multiple persons manually assess and label the qualities of *all* face images in a large database, this approach is desirable in that it only requires a small set of pairwise quality labels from each human rater in order to infer the quality ratings for the entire database. The details of data collection and the matrix completion approach are discussed in the remainder of this section.

1) Crowdsourcing Comparisons of Face Quality: Amazon Mechanical Turk (MTurk)[5] was utilized to facilitate collection of pairwise comparisons of face image quality from multiple human raters (*i.e.,* MTurk "workers"). Given a pair of face images, displayed side by side, our Human Intelligence Task (HIT) was to respond to the prompt "Indicate which face has better quality" by selecting one of the following: (i) left face is much better, (ii) left face is slightly better, (iii) both faces are similar, (iv) right face is slightly better, and (v) right face is much better. Fig. 5 shows the interface used to collect the responses.[6]

Our HIT requested each worker to provide responses to a total of 1,001 face image pairs from the LFW database, made up of 6 tutorial pairs, 974 random pairs, and 21 consistency check pairs. The tutorial pairs were pre-selected where the quality of one image was clearly better than the quality of the other (see Fig. 6). Because we expected these easy pairs to elicit "correct" responses, they allowed us to ensure that the worker had completed the tutorial introduction and understood the goal of the task. The next 974 pairs of images were chosen randomly from the LFW database, while the final 21 pairs were selected from the set of 974 as duplicates to test the consistency of the worker's responses. MTurk workers who attempted our HIT were only allowed to complete it if they passed the tutorial pairs, and we only accepted the submitted responses from workers who were consistent on at least 10 out of the 21 consistency check pairs.

Fig. 5. The interface used to collect pairwise comparisons of face quality from MTurk workers.

Fig. 6. Face images (from the LFW database) selected for the six tutorial pairs which are used to check whether MTurk workers understood the task before completing the pairwise comparisons used in our study of face image quality. Each of the tutorial pairs included one image from the top row (high face quality) and one from the bottom row (low face quality), so the face quality comparison was unambiguous.

In order to be eligible to attempt our HIT for assessment of face image quality, MTurk workers had to have previously completed at least 10,000 HITs from other MTurk "requesters" with an approval rating of at least 99%. These stringent qualifications helped to ensure that only experienced and reliable workers (in terms of MTurk standards) participated in our data collection.[7] A total of 435 MTurk workers began our HIT. After removing 245 workers who did not complete the full set of 1,001 pairwise comparisons and 4 workers who failed the consistency check (inconsistent response for 10 or more of the 21 duplicate pairs), a total of 194 workers were each compensated US $5.00 through the MTurk crowdsourcing service. In total, this quality labeling costed less than US $1,000 and all HITs were completed in less than one day.[8]

2) Matrix Completion: After collecting random sets of pairwise comparisons of face quality from 194 workers via MTurk, we use the matrix completion approach proposed by Yi *et al.* [27] to infer a complete set of quality ratings for each worker on the entire LFW database (13,233 total face

[5]https://www.mturk.com

[6]The tool is available at http://cse.msu.edu/~bestrow1/FaceOFF/.

[7]The MTurk worker qualifications are managed by the MTurk website.

[8]The 194,194 human assessments of pairwise face quality (collected for this study via Amazon Mechanical Turk) are publicly available for download as multimedia associated with this manuscript.

images). The aim is to infer $\hat{F} \in \mathbb{R}^{n \times m}$, the worker-rating matrix for face quality, where n is the number of workers and m is the number of face images. This framework is similar to the well-known Netflix recommendation system using matrix completion to infer user ratings from partial observations.

The matrix completion approach is based on a low-rank assumption for the matrix of interest, which, in our case, comes from the intuition that human assessments of face quality are unlikely to be totally independent from each other. Humans will give face quality ratings based on the same prominent factors of facial pose, illumination, sharpness, occlusion, etc. Using the low-rank assumption, Yi *et al.* [27] show that only $O(r \log m)$ pairwise queries are needed to infer the full ranking list of a worker for all m items (face images), where r is the rank of the unknown rating matrix ($r \ll m$). The maximum possible rank of the unknown rating matrix, $\hat{F}$, is $r = n = 194$ (number of workers), so $O(194 \log 13,233) \approx 800$; hence, the 974 random pairs per worker collected in our study are sufficient to do the matrix completion, especially since we expect $r < n$ (*i.e.,* the quality ratings from the n workers are not all independent).[9]

After matrix completion, the worker-rating matrix $\hat{F}$ (with dimensions $194 \times 13,233$) contains the inferred face quality ratings for all 13,233 LFW face images for each of the 194 workers. In analyzing the variability in the inferred quality ratings, we noticed an inverse relationship (Fig. 7) between the number of pairs a worker marked "Similar" and the resulting *range* of that worker's inferred quality ratings, where *range* = (max quality rating) − (min quality rating). [10] Because of this observation, *min-max* normalization was performed to transform the quality ratings from all workers to the same range of $[0, 1]$. With the aim of obtaining a single quality rating for each face image in the LFW database, we simply take the *median* quality rating from all 194 workers to reduce the $194 \times 13,233$ matrix to a $1 \times 13,233$ vector of quality ratings (one per image in LFW).

B. Matcher Quality Values (MQV)

Target face quality values derived from similarity scores serve as an "oracle" for a quality measure that is highly correlated with recognition performance. For example, if the goal is to detect and remove low-quality face images to improve the FNMR, then face images should be removed from a database in the order of their genuine scores. Previous works on biometric quality for fingerprint [3], [51] and face [30] have defined target or "ground truth" quality values as a measure of the separation between the sample's genuine score and its impostor distribution when compared to a gallery of enrollment samples. A normalized comparison score for the jth query sample of subject i can be defined as,

$$z_{ij} = (s_{ij}^G - \mu_{ij}^I)/\sigma_{ij}^I, \quad (1)$$

[9]Specifically, we use Yi *et al.*'s Transductive Crowdranking algorithm for matrix completion [27].

[10]This bias is not due to the coarse levels of "much better" and "slightly better" because before matrix completion, we combined these two levels to simply left/right image is "better".

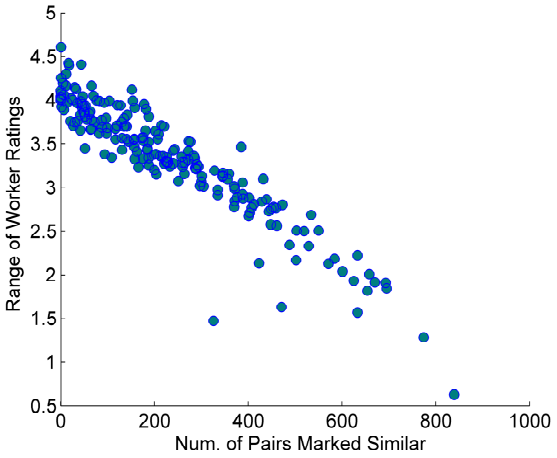


Fig. 7. The *range* of a worker's inferred face quality ratings (after matrix completion) inversely depends on the number of face image pairs that the worker marked "Similar" quality, where *range* = (max quality rating) − (min quality rating). Hence, *min-max* normalization was performed to transform the quality ratings from all workers to the same $[0, 1]$ range.

where s_{ij}^G is the genuine score and μ_{ij}^I and σ_{ij}^I are the mean and standard deviation, respectively, of the impostor scores for the query compared to the gallery. Previous works then bin the Z-normalized comparison scores into quality bins based on the cumulative distribution functions (CDFs) of sets of correctly and incorrectly matched samples [3], [30], [51]. Instead, we propose to directly predict z_{ij} for a given face image to obtain a continuous measure of face image quality.

Target quality values defined based on comparison scores are confounded by the fact that a comparison score is computed from *two* face images, but the aim is to label the quality of a *single* face image. Beveridge *et al.* [52] argue against face image quality as an intrinsic property of a single image by showing that face images can simultaneously be both easy and hard to recognize when compared with other images. Fig. 8 gives an example of this phenomenon.

To account for this conundrum, we make the simplifying assumption that similarity scores are generally governed by low-quality face images. While Fig. 8 shows an example of two poor-quality images matching with high genuine similarity, poor-quality images generally produce lower genuine similarity scores. Hence, face quality values derived from good/poor comparison scores can be assigned to probe images under the assumption that the quality of the enrollment images in a gallery are at least as good as the quality of probe images.

To establish a gallery of face images with quality assumed to be higher than a set of probe images, we manually selected the highest quality image available for each of the 1,680 subjects in the LFW database with at least two face images. Though this process unavoidably introduces some bias due to the authors' subjective judgments about face recognition covariates (pose, illumination, expression, etc.), there is no ground truth available to be used otherwise. Our goal is to establish this ground truth and evaluate it to determine if it useful. So, we use this manually-selected set of images as the gallery (1,680 images, one per subject), while the remaining 7,484 images of these subjects are used as the probe set. The

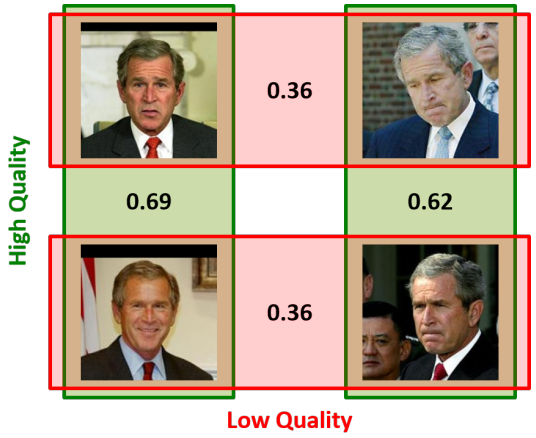


Fig. 8. Illustration of the pairwise quality issue. Face images in the left and right columns are individually of high and low quality, respectively. However, when compared, the images can produce both high (good) and low (bad) genuine similarity scores. Scores are from COTS-A with range of $[0, 1]$.

additional 4,069 images in the LFW database from subjects with only a single image are used to extend the size of the gallery. Normalized comparison scores, z_{ij}, are computed using Eqn. (1) for the 7,484 probe images for each of the face matchers (COTS-A, COTS-B, and ConvNet) and are used as target matcher quality values (MQV) for learning the matcher-dependent face quality predictor.

V. Automatic Prediction of Face Quality

Given that we have established target face quality values for the LFW database, we now wish to train a model to automatically predict the quality of other unconstrained face images. With the of influx of deep learning, over the last several years, almost all face recognition systems since proposed are also CNN-based. Because face quality should be a predictor of face recognition performance, we make use of the same features which most FR systems now use. ConvNet face recognition has been extremely successful for unconstrained images with representations already robust to a lot of quality factors which were always seen as confounding to performance. So, given a representation which has been trained for recognition purposes, we want to see if we can train a model to distinguish between the quality factors which have not been suppressed by training for recognition purposes.

The ConvNet used to extract features for the proposed face quality predictors is based on Wang *et al.*'s network for face recognition, which is similar to the networks in [19], [20]. The ConvNet is trained on the CASIA database [19] for face recognition purposes. Faces in RGB images are detected by the Dlib[11] face detector, the face is rotated upright base on the eye locations, and the mid-point between the leftmost and rightmost landmarks is used to horizontally center the face.[12] Images are resized to 110 $\times$ 110 pixels with the eyes and mouth placed 45% and 25% from the top and bottom of the image, respectively.[13] The architecture of the ConvNet includes 10 convolutional layers and 5 pooling layers (more details given in [17]), and outputs a 320-dimensional feature vector. We directly use this representation, which we refer to as *Deep-320* features, as the feature vector for predicting both MQV and HQV face quality.

Using the Deep-320 face image features, we train a support vector regression (SVR) [53] model with radial basis kernel function (RBF) to predict either the normalized comparison scores, z_{ij}, from face matchers (for MQV predictor) or the human quality ratings (for HQV predictor). The parameters for SVR (cost, epsilon, and gamma for RBF) are determined via grid search on a validation set of face images.

VI. Experimental Evaluation

The aim of this work is twofold: (i) establish the target, or "ground truth", quality values of a face image database, and (ii) use this quality-labeled face image database to train a model to predict the quality values using features automatically extracted from an unseen test face image (prior to matching). Hence, in Sec. VI-A, we first evaluate the *target* face quality values to determine their utility for automatic recognition. In Sec. VI-B1 we then evaluate how well the target quality values can be predicted by the proposed model for automatic face image quality on the LFW database, and in Sec. VI-B2 we evaluate the utility of the proposed face image quality values for recognition of face images and video frames from the IJB-A database [6].

Following the methodology advocated by Grother and Tabassi [3], we evaluate the face quality measures using the *Error versus Reject (EvR) curve* which evaluates the efficiency of rejecting low quality samples for reducing error rates. The EvR curve plots an error rate (FNMR or FMR) versus the fraction of images removed/rejected, where the error rates are re-computed using a fixed threshold (*e.g.,* overall FMR = 0.01%) after a fraction of the images have been removed. We additionally evaluate the utility of the proposed face image quality predictors for improving template-based matching in the IJB-A protocol [6] and provide visual inspections of face images rank-ordered by the proposed face quality predictors.

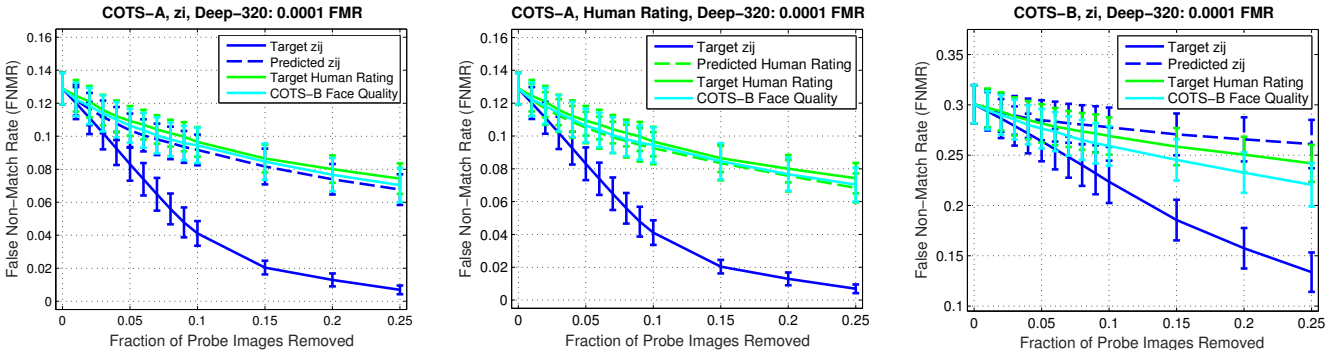
A. Target Face Quality Values

Using the gallery and probe setup of the LFW database detailed in Section IV-B, Fig. 10 plots EvR curves for removing probe images based on MQV or HQV target (ground-truth) face quality values to reduce FNMR of the three face matchers (COTS-A, COTS-B, and ConvNet). Because the matchers are of different strengths, a common initial FNMR and FMR of 0.20 and 0.10, respectively, were chosen for the evaluation of all three matchers. Fig. 10a shows that removing low-quality probe images in order of HQV decreases FNMR for all three matchers, indicating that human quality ratings are correlated with recognition performance. However, MQV is much more efficient in reducing FNMR. This is expected because the

[11] http://blog.dlib.net/2014/08/real-time-face-pose-estimation.html

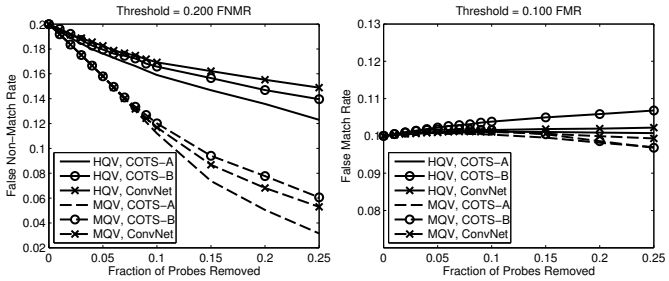
[12] If Dlib landmark detection fails, the ground truth landmarks provided with the IJB-A database are used instead (typically for extreme profile faces).

[13] Whenever COTS-A or COTS-B are used, the raw face images are input to the matcher, so the respective matcher conducts its own face detection, alignment, and cropping.



(a) COTS-A, z_{ij} (b) COTS-A, Human (c) COTS-B, z_{ij}

Fig. 9. Error vs. Reject curves for target and predicted face image quality values (MQV and HQV) for the LFW database. The curves show the efficiency of rejecting low quality face images in reducing FNMR at a fixed FMR of 0.01%. The models used for the face quality predictions in (a)-(c) are SVR on the deep-320 features from ConvNet in [17].



(a) FNMR (b) FMR

Fig. 10. Error vs. Reject curves for (a) FNMR and (b) FMR on the LFW database (gallery size of 5,749 face images and 7,484 probe face images from LFW [26]). Probe images were rejected in order of *target* human quality values (HQV) or matcher quality values (MQV). Thresholds are fixed at (a) 0.20 FNMR and (b) 0.10 FMR for comparison of the three face matchers (COTS-A, COTS-B, and ConvNet [17]).

score-based target quality values are computed from the same comparison scores used to compute the FNMR and serve as an "oracle" for a desirable quality measure.

The utility of the target quality values in terms of reducing FMR in Fig. 10b is not as apparent; in fact, removing low-quality images based on HQV clearly *increases* FMR for COTS-B, though the magnitude of the increase is small (removing 25% of the probe images increases FMR by 0.14%). The relation between face quality and impostor scores (*i.e.,* FMR) is generally less of a concern. For biometric quality, in general, we desire *high* quality samples to produce *low* impostor similarity scores, but *low* quality samples may also produce *low* (or even lower) impostor scores. If this is the case, low quality face images may be beneficial to FMR for empirical evaluation, but still undesirable operationally. Due to this conundrum, we focus on the effect of face quality on FNMR for the remainder of the experiments.

B. Predicted Face Quality Values

The proposed framework for automatic prediction of face image quality (using both human ratings and score-based quality values as targets) is used to predict the quality of face images from the LFW [26] and IJB-A [6] databases. The

TABLE III
SPEARMAN RANK CORRELATION (MEAN ± STANDARD DEVIATION OVER 10 RANDOM SPLITS OF LFW IMAGES) BETWEEN TARGET AND PREDICTED MQV AND HQV

	Matcher		
	COTS-A	COTS-B	ConvNet
MQV	0.558 ± 0.023	0.442 ± 0.026	0.459 ± 0.022
HQV	0.585 ± 0.019		

prediction models for both databases are trained using LFW face images, and the experimental protocols are detailed in the following sections.

1) Train, Validate, and Test on LFW: We first divide 7,484 face images of the 1,680 subjects with two or more images in LFW into 10 random splits for training and testing, where $2/3$ and $1/3$ of the subjects are randomly split into training and testing sets, respectively. For each split, we then conduct 5-fold cross-validation within the training set to tune the parameters for the SVR model via grid search. The selected set of parameters is applied to the full training set to result in a single model for each of the 10 splits, which are then used to predict the quality labels of the images in each of the 10 test sets. This framework ensures subject-disjoint training and testing sets, and parameter selection is conducted within a validation set, not optimized for the test sets.

Table III gives the rank correlation (mean and standard deviation over the 10 splits) between the target and predicted quality values for HQV and MQV (MQV separately for COTS-A, COTS-B, and ConvNet matchers). We observe that prediction of HQV is more accurate than prediction of MQV for all three matchers, likely due to the difficulty in predicting particular nuances of each matcher. Fig. 11 shows images sorted by predicted HQV of four example subjects from LFW with *strong* rank correlation (Spearman) between target and predicted human quality values.

To evaluate the quality values in the context of automatic face recognition performance, EvR curves (for FNMR at fixed 0.01% FMR) are plotted in Fig. 9 for both target and predicted quality values (MQV and HQV). The figures demonstrate that rejecting low quality face images based on predicted z_{ij},

Fig. 11. Face images of four subjects from LFW [26] rank-ordered by the predicted human quality ratings from the proposed HQV method. Face images are shown in order of decreasing face quality.

predicted HQV, or the COTS-B measure of face quality, results in comparable efficiency in reducing FNMR (*e.g.,* removal of 5% of probe images lowers FNMR by ∼2%). However, none of the methods are near as efficient as rejecting images based on the target z_{ij} values, which serve as an oracle for a predicted face quality measure that is highly correlated with the recognition performance.

2) Train and Validate on LFW, Test on IJB-A: In this framework, we conduct 5-fold cross-validation over the 7,484 LFW images (folds are subject-disjoint) to determine the parameters for the SVR model via grid search. We then apply the selected set of parameters to all of the LFW training images. This model trained on the LFW database is then used to predict the quality of face images in the IJB-A database [6].

For evaluation on the IJB-A database, we follow the template-based verification protocol [6], which consists of 10 random splits (bootstrap samples) of the 500 total IJB-A subjects. For each split, 333 subjects are randomly sampled for training and the remaining 167 subjects for testing. Note that we do not do any fine-tune training with IJB-A images; our face quality models are trained using only the LFW and CASIA databases. In template-based matching, multiple face images and/or video frames are available for a subject in the gallery and/or probe sets. Baseline results for score-level fusion (SLF) using *max* and *mean* rules are given in Figure 13a for the COTS-A and ConvNet matchers. COTS-B was not used for evaluation on IJB-A database because of a much higher failure to enroll (FTE) rate than COTS-A and ConvNet matchers. Figure 13a shows that mean fusion is slightly better than max fusion for both matchers and that at 1% FMR, COTS-A and ConvNet are comparable.

For the IJB-A database, we compare the proposed MQV and HQV methods with Chen *et al.*'s Rank-based Quality Score (RQS) [35]. The RQS method defines pairwise constraints on face images based on a relative ordering of face image databases. The learning to rank (L2R) framework of Parikh and Grauman [44] is used to learn five different ranking functions, one for each of five different image features (HoG, Gabor, Gist, LBP, and CNN features), which we refer to as *Feat-5*. The five ranking functions are then combined with a polynomial kernel mapping (PKM) function. To predict the RQS of a new test image, Feat-5 image features are extracted and multiplied by the weight vector obtained from the (L2R + PKM) framework.

Using the RQS[14] and the L2R[15] codes, both made publicly available by their authors, we combine different components of the RQS method with the human pairwise comparisons from MTurk and the Deep-320 features to evaluate the impact of these components. Figure 12 shows a flowchart of the variants of the proposed HQV method (HQV-0, HQV-1, and HQV-2), where MTurk pairwise comparisons are the input to establish target quality values, but Feat-5 features and/or (L2R + PKM) framework are used instead of Deep-320 features and/or matrix completion. The flowcharts for the proposed MQV method and Chen *et al.*'s RQS method are also given in Fig. 12; in total, we evaluate five different face quality methods.

Finally, to evaluate the utility of face quality values to recognition performance, we incorporate the face quality into template-based matching as follows: given a threshold on the face quality, the template for a subject consists of only the faces with quality at least as high as the threshold; if there are no faces with quality above the threshold, select only the single best face. Score-level mean fusion is then applied to the scores from the selected faces. Note that COTS-A is a still image face matcher; we apply COTS-A individually to all pairs of faces in two templates and then do score-level fusion. Hence, COTS-A is not also doing anything internally to detect/reject low-quality images for template-based matching.

Figures 13b and 13c report the reduction in FNMR at fixed 1% FMR when the threshold on face quality is varied; the thresholds considered are $n/100$ where n is the nth percentile of the face quality values for all images and videos in the given testing split of IJB-A database. This evaluation is similar to the EvR curve except that the number of scores used to compute performance remains the same as face samples are removed from the templates. Because the face quality methods

[14]http://jschenthu.weebly.com/projects.html
[15]https://filebox.ece.vt.edu/~parikh/relative.html

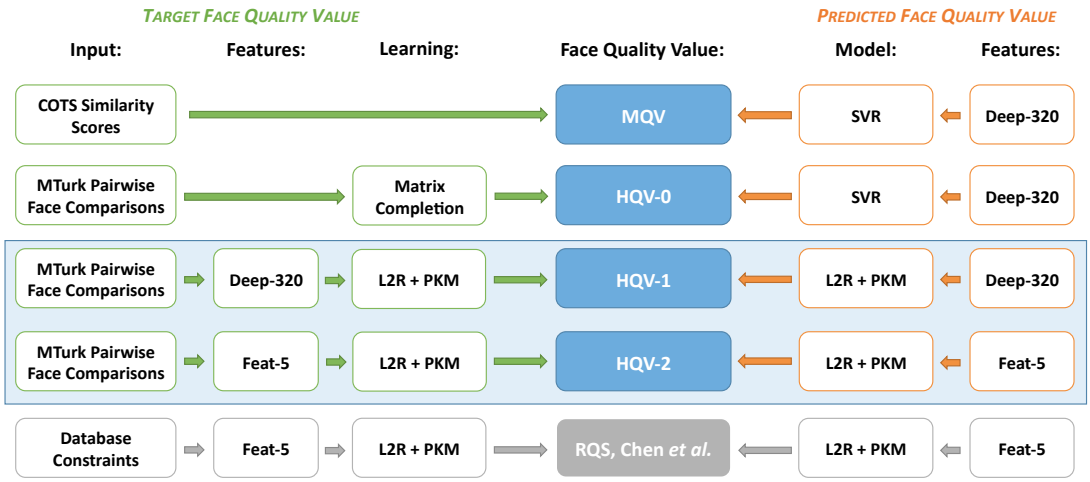
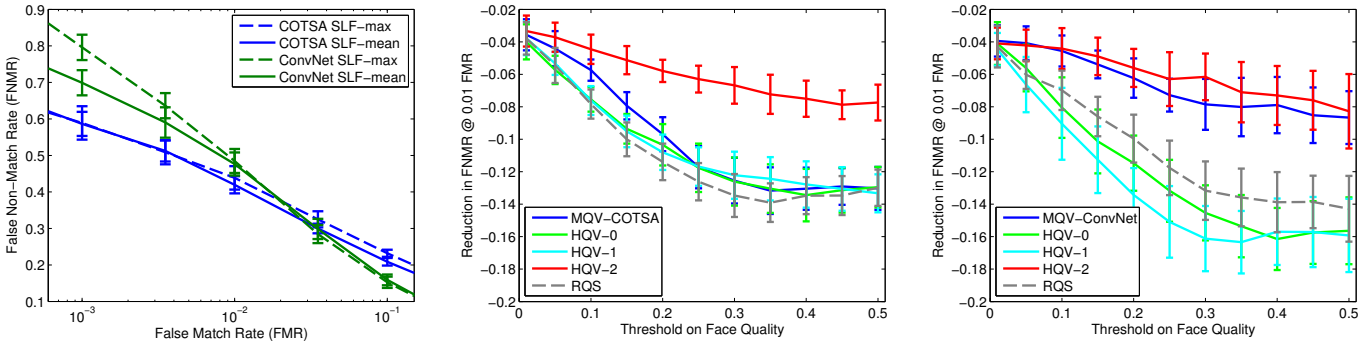


Fig. 12. Flowcharts indicating the components of defining target face quality values (left side) and then predicting the target face quality values (right side) for the five methods evaluated on the IJB-A database [6]. MQV and HQV-0 are the methods proposed in this paper, while RQS is proposed by Chen *et al.* [35]. HQV-1 and HQV-2 are variants of HQV-0 with some components of RQS plugged in to evaluate the impact of the input used to define target quality values, the image features (Deep-320 vs. Feat-5), and the learning framework used (support vector regression (SVR) vs. learning to rank with polynomial kernel mapping function (L2R + PKM)).



(a) COTS-A and ConvNet (b) COTS-A (c) ConvNet

Fig. 13. Results for the verification protocol of the IJB-A database [6]. All curves in (a)-(c) show mean performance and error bars give standard deviation in performance over the 10 splits in the protocol. (a) Receiver Operating Characteristic (ROC) for COTS-A and ConvNet [17] matchers, where score-level fusion (SLF) is applied to the multiple face samples per subject for template-based matching of the IJB-A protocol. Using thresholds on face image quality measures to determine which face samples in a template to use for matching, (b) and (c) plot reduction in FNMR at 1% FMR, showing that FNMR decreases as the face quality thresholds are increased. Flowcharts providing details of each face quality method (MQV, HQV-0, etc.) are given in Fig. 12. The RQS method is proposed by Chen *et al.* [35].

MQV, HQV, and RQS each use their own face detection and alignment methods, the face quality for images in which *any* of the detection/alignment failed (*i.e.,* failed to enroll) are all set to the lowest quality value, so these images are removed first for all face quality methods, providing a fairer comparison.

A few observations can be made from Figs. 13b and 13c about the different face quality methods. (i) MQV performs quite well at reducing FNMR for COTS-A, but is much worse for the ConvNet matcher. This may be because the Deep-320 features used for MQV face quality prediction are the same features used by the ConvNet matcher, so the MQV for ConvNet was trained on the same similarity scores and image features that the quality is trying to predict. (ii) HQV-2 performs poorly, while HQV-1 effectively reduces FNMR for both matchers, suggesting that Deep-320 features are more powerful for predicting the human quality ratings than Feat-5 features. (iii) HQV-0 and HQV-1 perform comparably for COTS-A, but HQV-1 performs slightly better for the ConvNet matcher. This suggests that the (L2R + PKM) framework may be somewhat better than matrix completion for establishing the target face quality values from pairwise comparisons. (iv) HQV-0 and HQV-1 both perform comparable to the RQS method [35] for both matchers, and all three face quality methods effectively reduce FNMR by removing low-quality face images or videos from IJB-A templates. Using mean score-level fusion of all faces in the templates as a baseline, FNMR is reduced by $\sim$13% for COTS-A and $\sim$16% for ConvNet matchers given a quality threshold of the 40th percentile of the distribution of quality values in the training sets.

Table IV summarizes the results on the IJB-A verify protocol [6] for the COTS-A and ConvNet matchers with and without the proposed HQV face quality predictor and compares the performance to previously published results on IJB-A. Performance is reported as TAR at fixed FARs as the protocol

TABLE IV
VERIFICATION PERFORMANCE FOR THE IJB-A DATABASE [6]. RESULTS ARE REPORTED AS AVERAGE ± STANDARD DEVIATION OVER THE 10 FOLDS IN THE IJB-A VERIFY PROTOCOL.

Algorithm	Ref.	TAR (%) @ 0.1% FAR	TAR (%) @ 1.0% FAR
CNN+AvgPool (baseline)	[43]	77.1 ± 6.4	91.3 ± 1.4
NAN	[43]	88.1 ± 1.1	94.1 ± 0.8
TPE (baseline)	[49]	71.3 ± 5.0	85.4 ± 1.0
TPE	[49]	81.3 ± 2.0	90.0 ± 1.0
Masi *et al.* (baseline)	[46]	53.0 ± *n/a*	79.9 ± *n/a*
Masi *et al.*	[46]	72.5 ± *n/a*	88.6 ± *n/a*
Rapid Synthesis (baseline)	[47]	68.0 ± *n/a*	85.2 ± *n/a*
Rapid Synthesis	[47]	75.0 ± *n/a*	88.8 ± *n/a*
DR-GAN (baseline)	[21]	57.7 ± 3.7	77.3 ± 1.9
DR-GAN	[21]	69.9 ± 2.9	83.1 ± 1.7
$\text{PAMs}_{frontal}$ (baseline)	[48]	55.2 ± 3.2	73.3 ± 1.8
PAMs	[48]	65.2 ± 3.7	82.6 ± 1.8
DCNN (baseline)	[20]	*n/a*	57.3 ± 2.4
$\text{DCNN}_{ft+metric}$	[20]	*n/a*	78.7 ± 4.3
DCNN_{fusion}	[20]	*n/a*	83.8 ± 4.2
ConvNet, # nets = 1 (baseline)	ours	30.1 ± 3.5	52.3 ± 3.2
ConvNet, # nets = 1, w/ HQV	ours	48.0 ± 5.1	68.5 ± 3.5
ConvNet, # nets = 6	[17]	51.4 ± 6.0	73.3 ± 3.4
COTS-A (baseline)	ours	41.4 ± 3.3	58.0 ± 2.4
COTS-A w/ HQV	ours	61.7 ± 2.7	71.5 ± 1.3

Note: *n/a* indicates performance was not reported

suggests [6]. In addition to the state-of-the-art methods on IJB-A, we have also included each method's baseline performance previously reported in "ablation" studies provided by the original authors to demonstrate the improvements attributed to different components of their proposed end-to-end recognition algorithms. Masi *et al.*'s group uses the VGGNet architecture fine-tuned on the CASIA database, but achieve significant performance improvements by augmenting the training and testing sets with synthetically generated faces at different poses and expression [46]. The ConvNet architectures used in our work and by [20] are both based on the architecture originally proposed in [19], but [20] uses parametric rectified linear unit (PReLu) instead of the rectified linear unit (ReLu) and fuse the similarity scores from two networks trained on gray-scale and RGB images.

Table IV shows that our baselines of COTS-A and ConvNet matchers are initially poor compared with the baseline architectures of the other methods. However, simply using the proposed HQV to reject poor quality face images prior to score-level fusion greatly improves the performance of both COTS-A and ConvNet. At 0.1% FAR, TAR increases by 17.9% for COTS-A and 20.3% for ConvNet matchers. This improvement is on the same order of magnitude or larger than the improvement achieved by the other methods over their baselines, except for [20]. However, 21.4% of the improvement in TAR at 1% FAR for DCNN method is due to fine-tuning on the IJB-A training sets with Joint Bayesian metric learning [20]. Recall that we do not do any fine-tuning on the IJB-A training sets and simply use score-level (mean) fusion as the similarity measure.

Figure 14 shows examples of face images (and video frames in Fig. 15) sorted in order of the proposed automatic face quality prediction for human quality ratings (HQV-0). Fig. 14 also shows face images sorted by RQS *et al.* [35] for comparison. Visually, both methods appear to do reasonably well at ranking face images by quality, where both methods are noticeably sensitive to facial pose, in particular.

VII. CONCLUSIONS

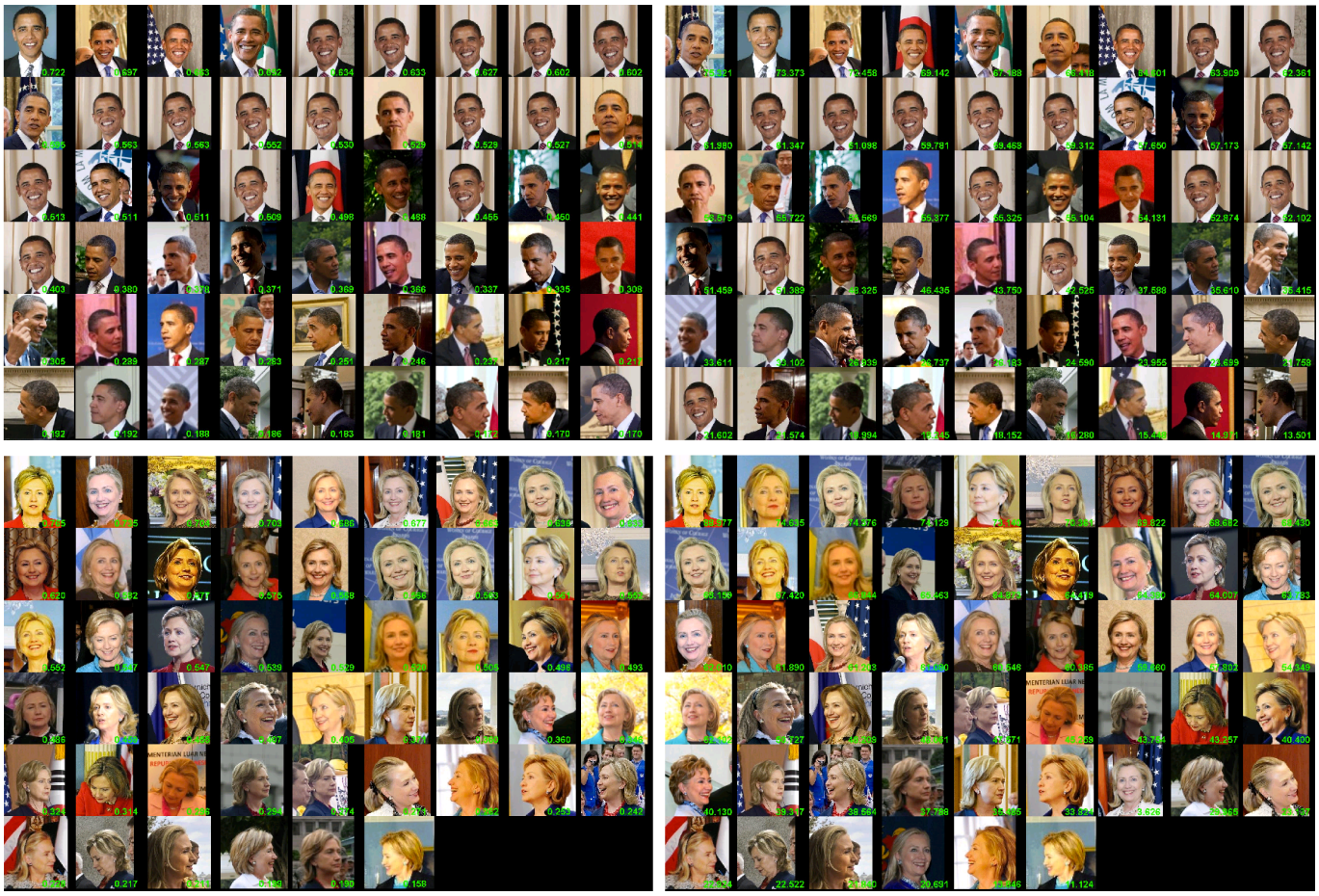
Automatic face quality assessment is a challenging problem with important operational applications. Automatic detection of low-quality face images would be beneficial for maintaining the integrity of enrollment databases, reacquisition prompts, quality-based fusion, and adaptive recognition approaches. In this work, we proposed a model for automatic prediction of face quality using image features extracted prior to matching. The conclusions and contributions are summarized as follows:

- Human ratings of face quality are correlated with recognition performance for unconstrained face images. Rejection of 5% of the lowest quality faces (based on the proposed HQV) in the LFW database resulted in $\sim$2% reduction in FNMR, while using HQV to select subsets of images for template-based matching of IJB-A database reduced FNMR by at least 13% (at 1% FMR) for two different matchers (COTS-A and ConvNet).
- Automatic prediction of human quality ratings (HQV) is more accurate than prediction of score-based face quality values (MQV). MQV prediction is challenging because of nuances of specific matchers and pairwise quality factors (*i.e.,* comparison scores are a function of *two* faces, but the scores are used to label the quality of a *single* face).
- The proposed HQV method performs comparably to Chen *et al.*'s RQS [35] for quality-based selection when multiple face images/videos are available for a subject.
- Visual inspection of face images rank-ordered by the proposed face quality measures (both HQV and MQV) are promising, even for cross-database prediction (*i.e.,* model trained on LFW [26] and tested on IJB-A [6]).

The image features used in this work extracted from a deep ConvNet, which was trained for recognition purposes, show promising results for face image quality prediction. However, this face representation should ideally be robust to face quality factors. It would be interesting to retrain a ConvNet for prediction of face image quality, rather than recognition, to compare which of these two methods for training a quality predictor is better for a face quality measure that is correlated with recognition performance.

REFERENCES

[1] P. J. Grother, G. W. Quinn, and P. J. Phillips, "Report on the evaluation of 2D still-image face recognition algorithms," NIST Interagency Report 7709, Aug. 2011.
[2] P. Grother and M. Ngan, "FRVT: Performance of face identification algorithms," NIST Interagency Report 8009, May 2014.
[3] P. Grother and E. Tabassi, "Performance of biometric quality measures," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 531–543, Apr. 2007.
[4] S. Bharadwaj, M. Vatsa, and R. Singh, "Biometric quality: A review of fingerprint, iris, and face," *EURASIP Journal on Image and Video Processing*, vol. 34, no. 1, 2014.
[5] F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia, "Quality measures in biometric systems," *IEEE Security Privacy*, vol. 10, no. 6, pp. 52–62, Nov. 2012.



(a) Ranked by the Proposed HQV (b) Ranked by RQS [35]

Fig. 14. Face images from two subjects in IJB-A [6] sorted by face quality (best to worst). The face quality was predicted by (a) the proposed HQV predictor (SVR model to predict human quality values from Deep-320 features [17]) and (b) Rank-based Quality Score (RQS) [35] for comparison.

Fig. 15. Cropped faces from the videos of three subjects in IJB-A [6] sorted by face image quality (best to worst) automatically predicted by the proposed HQV method (SVR on Deep-320 image features [17] trained on human quality ratings from the LFW database).

[6] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A," in *Proc. CVPR*, 2015.

[7] N. Poh and J. Kittler, "A unified framework for biometric expert fusion incorporating quality measures," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 3–18, Jan 2012.

[8] N. Poh, T. Bourlai, J. Kittler, L. Allano, F. Alonso-Fernandez, O. Ambekar, J. Baker, B. Dorizzi, O. Fatukasi, J. Fierrez, H. Ganster, J. Ortega-Garcia, D. Maurer, A. A. Salah, T. Scheidat, and C. Vielhauer, "Benchmarking quality-dependent and cost-sensitive score-level multimodal biometric fusion algorithms," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 4, pp. 849–866, Dec 2009.

[9] F. Alonso-Fernandez, J. Fierrez, D. Ramos, and J. Gonzalez-Rodriguez, "Quality-based conditional processing in multi-biometrics: Application to sensor interoperability," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 40, no. 6, pp. 1168–1179, Nov 2010.

[10] J. Roth, Y. Tong, and X. Liu, "Adaptive 3d face reconstruction from unconstrained photo collections," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Dec. 2016.

[11] B. F. Klare, M. Burge, J. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Trans. on Information Forensics and Security*, vol. 7, no. 6, pp. 1789–1801, Dec. 2012.

[12] P. J. Phillips, M. Q. Hill, J. A. Swindle, and A. J. O'Toole, "Human and algorithm performance on the PaSC face recognition challenge," in *Proc. BTAS*, 2015.

[13] W. J. Scheirer, P. J. Flynn, C. Ding, G. Guo, V. Struc, M. A. Jazaery, K. Grm, S. Dobrisek, D. Tao, Y. Zhu, J. Brogan, S. Banerjee, A. Bharati, and B. R. Webster, "Report on the BTAS 2016 video person recognition evaluation," in *IEEE Conf. on Biometrics Theory, Applications and Systems*, 2016.

[14] A. Blanton, K. C. Allen, T. Miller, N. D. Kalka, and A. K. Jain, "A comparison of human and automated face verification accuracy on

unconstrained image sets," in *Computer Vision and Pattern Recognition, Workshop on Biometrics*, 2016.
[15] A. Adler and T. Dembinsky, "Human vs. automatic measurement of biometric sample quality," in *Canadian Conf. on Electrical and Computer Engineering*, 2006.
[16] R.-L. Hsu, J. Shah, and B. Martin, "Quality assessment of facial images," in *Biometrics Symposium: Special Issue on Research at the Biometric Consortium Conf.*, 2006.
[17] D. Wang, C. Otto, and A. K. Jain, "Face search at scale: 80 million gallery," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, Jun. 2016.
[18] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Computer Vision and Pattern Recognition*, 2013.
[19] D. Yi, S. Liao, and S. Z. Li, "Learning face representation from scratch," arXiv:1411.7923, Nov. 2014.
[20] L.-C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep CNN features," in *Winter Conf. on Applications of Computer Vision*, 2016.
[21] L. Tran, X. Yin, and X. Liu, "Representation learning by rotating your faces," *CoRR*, vol. abs/1705.11136, 2017. [Online]. Available: http://arxiv.org/abs/1705.11136
[22] J. R. Beveridge, H. Zhang, B. A. Draper, P. J. Flynn, Z. Feng, P. Huber, J. Kittler, Z. Huang, S. Li, Y. Li, M. Kan, R. Wang, S. Shan, X. Chen, H. Li, G. Hua, V. truc, J. Kriaj, C. Ding, D. Tao, and P. J. Phillips, "Report on the FG 2015 video person recognition evaluation," in *IEEE Conf. on Automatic Face and Gesture Recognition*, 2015.
[23] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2005.
[24] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer, "The good, the bad, and the ugly face challenge problem," *Image and Vision Computing*, vol. 30, no. 3, pp. 177–185, 2012.
[25] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, May 2010.
[26] 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," U. of Amherst, Tech. Report 07-49, Oct. 2007.
[27] J. Yi, R. Jin, S. Jain, and A. K. Jain, "Inferring users' preferences from crowdsourced pairwise comparisons: A matrix completion approach," *First AAAI Conf. on Human Computation and Crowdsourcing*, 2013.
[28] G. Aggarwal, S. Biswas, P. J. Flynn, and K. W. Bowyer, "Predicting performance of face recognition systems: An image characterization approach," in *IEEE CVPRW*, 2011.
[29] P. J. Phillips, J. R. Beveridge, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, S. Cheng, M. N. Teli, and H. Zhang, "On the existence of face quality measures," in *IEEE Conf. on Biometrics: Theory, Applications and Systems*, Sep. 2013, pp. 1–8.
[30] S. Bharadwaj, M. Vatsa, and R. Singh, "Can holistic representations be used for face biometric quality assessment?" in *IEEE Int'l Conf. on Image Processing*, 2013.
[31] A. Abaza, M. A. Harrison, T. Bourlai, and A. Ross, "Design and evaluation of photometric image quality measures for effective face recognition," *IET Biometrics*, vol. 3, no. 4, pp. 314–324, Dec. 2014.
[32] A. Dutta, R. Veldhuis, and L. Spreeuwers, "A bayesian model for predicting face recognition performance using image quality," in *IEEE Int'l Joint Conf. on Biometrics*, 2014.
[33] H. I. Kim, S. H. Lee, and Y. M. Ro, "Face image assessment learned with objective and relative face image qualities for improved face recognition," in *IEEE Int'l Conf. on Image Processing*, Sep. 2015.
[34] S. Vignesh, K. M. Priya, and S. S. Channappayya, "Face image quality assessment for face selection in surveillance video using convolutional neural networks," in *2015 IEEE Global Conf. on Signal and Information Processing (GlobalSIP)*, 2015.
[35] J. Chen, Y. Deng, G. Bai, and G. Su, "Face image quality assessment based on learning to rank," *IEEE Signal Processing Letters*, vol. 22, no. 1, pp. 90–94, 2015.
[36] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, D. S. Bolme, and Y. M. Lui, "FRVT 2006: Quo vadis face quality," *Image and Vision Computing*, vol. 28, no. 5, pp. 732–743, May 2010.
[37] J. Beveridge, G. Givens, P. J. Phillips, and B. Draper, "Factors that influence algorithm performance in the face recognition grand challenge," *Comput. Vis. Image Underst.*, vol. 113, no. 6, pp. 750–762, Jun. 2009.
[38] H. Sellahewa and S. A. Jassim, "Image-quality-based adaptive face recognition," *IEEE Trans. on Instrumentation and Measurement*, vol. 59, no. 4, pp. 805–813, Apr. 2010.
[39] L. Best-Rowden, H. Han, C. Otto, B. Klare, and A. K. Jain, "Unconstrained face recognition: Identifying a person of interest from a media collection," *IEEE Trans. on Information Forensics and Security*, vol. 9, no. 12, pp. 2144–2157, Dec. 2014.
[40] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *IEEE CVPRW*, Jun. 2011, pp. 74–81.
[41] G. Goswami, R. Bhardwaj, R. Singh, and M. Vatsa, "Mdlface: Memorability augmented deep learning for video face recognition," in *IEEE Int'l Joint Conf. on Biometrics*, Sept 2014, pp. 1–7.
[42] G. Goswami, M. Vatsa, and R. Singh, "Face verification via learned representation on feature-rich video frames," *IEEE Trans. on Information Forensics and Security*, vol. 12, no. 7, pp. 1686–1698, Jul. 2017.
[43] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017.
[44] D. Parikh and K. Grauman, "Relative attributes," in *Proc. Int'l Conf. on Computer Vision*, 2011.
[45] P. Viola and M. J. Jones, "Robust real-time face detection," *Int J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.
[46] I. Masi, A. Tran, T. Hassner, J. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *European Conf. on Computer Vision*, 2016.
[47] I. Masi, T. Hassner, A. T. Tran, and G. Medioni, "Rapid synthesis of massive face sets for improved face recognition," in *IEEE Conf. on Automatic Face and Gesture Recognition*, 2017.
[48] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016.
[49] S. Sankaranarayanan, A. Alavi, C. D.Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in *IEEE Conf. on Biometrics, Theory, Applications and Systems*, 2016.
[50] S. Liao, Z. Lei, D. Yi, and S. Z. Li, "A benchmark study on large-scale unconstrained face recognition," in *Int'l Joint Conf. on Biometrics*, 2014.
[51] E. Tabassi and C. L. Wilson, "A novel approach to fingerprint image quality," in *IEEE Int'l Conf. on Image Processing*, 2005.
[52] J. R. Beveridge, P. J. Phillips, G. H. Givens, B. A. Draper, M. N. Teli, and D. S. Bolme, "When high-quality face images match poorly," in *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, 2011.
[53] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Lacey Best-Rowden received the B.S. degree in computer science and mathematics from Alma College, Alma, Michigan, in 2010, and the Ph.D. from the Dept. of Computer Science and Engineering at Michigan State University, East Lansing, Michigan, in 2017. She is currently a Senior Biometrics Researcher at Visa Inc., Foster City, CA, USA. Her research interests include pattern recognition, computer vision, and image processing with applications in biometrics.

Anil K. Jain is a University distinguished professor in the Dept. of Computer Science and Engineering at Michigan State University. He is a Fellow of the ACM, IEEE, IAPR, AAAS and SPIE. 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, a member of the United States Defense Science Board and the Forensics Science Standards Board. He has received Fulbright, Guggenheim, Alexander von Humboldt, and IAPR King Sun Fu awards. He is a member of the United States National Academy of Engineering and a Foreign Fellow of the Indian National Academy of Engineering.