This is a pre-print of the original paper accepted for oral presentation at the International Joint Conference on Biometrics (IJCB) 2017.

Unconstrained Face Detection and Open-Set Face Recognition Challenge

M. Günther,[a] P. Hu,[b] C. Herrmann,[c] C. H. Chan,[d] M. Jiang,[e] S. Yang,[f] A. R. Dhamija,[a] D. Ramanan,[b]
J. Beyerer,[c] J. Kittler,[d] M. Al Jazaery,[e] M. I. Nouyed,[e] G. Guo,[e] C. Stankiewicz,[f] and T. E. Boult[a]
[a]University of Colorado Colorado Springs, [b]Carnegie Mellon University Pittsburgh, [c]Karlsruhe Institute of Technology,
[d]University of Surrey, [e]West Virginia University, [f]University of Wolverhampton

Abstract

Face detection and recognition benchmarks have shifted toward more difficult environments. The challenge presented in this paper addresses the next step in the direction of automatic detection and identification of people from outdoor surveillance cameras. While face detection has shown remarkable success in images collected from the web, surveillance cameras include more diverse occlusions, poses, weather conditions and image blur. Although face verification or closed-set face identification have surpassed human capabilities on some datasets, open-set identification is much more complex as it needs to reject both unknown identities and false accepts from the face detector. We show that unconstrained face detection can approach high detection rates albeit with moderate false accept rates. By contrast, open-set face recognition is currently weak and requires much more attention.

1. Introduction

Automatic face recognition is an important field and has a tremendous impact on many domains of our life. For example, private images can be sorted by persons that appear on them (e.g., Apple Photos or Google Photos), or airports perform automatic face recognition as passport control [29]. As the latter has severe security implications, most face recognition challenges such as the Face Recognition Vendor Tests[1] evaluate algorithms that perform verification, i.e., where a pair of model and probe images is tested whether they contain the same identity. Usually, a similarity between model and probe image is thresholded, where the threshold is computed based on a desired false acceptance rate. Other challenges included more difficult data, such as the Point and Shoot Challenge [6] or the Face Recognition Evaluation in Mobile Environment [11].

On the other hand, identification seems to be a more intricate problem, as a probe image must be compared to all identities enrolled in a gallery. As Klontz and Jain [20] and Kemelmacher-Shlizerman *et al.* [18] showed, when the gallery is large and probe images are taken in uncontrolled conditions, identifying the correct person is not trivial. In real surveillance scenarios subjects usually do not realize that their faces are captured and, hence, do not cooperate with the system. Furthermore, most of the captured faces will not belong to any person in the gallery and should be declared as unknown, leading to open-set face recognition. Also, face detectors might have false accepts, i.e., where a region of the background is detected as a face. These misdetections also need to be classified as unknown by face recognition algorithms. Therefore, additionally to identifying the correct person in the gallery based on difficult imagery, for an unknown face or misdetection, the similarity to *all* persons in the gallery must be below a certain threshold, which is usually computed based on a desired false identification rate. While the latest face recognition benchmark IJB-A [19] includes an open-set protocol, it does not treat misdetections that are subsequently labeled with an identity as an error, which makes that benchmark incomplete.

For the UCCS unconstrained face detection and open-set face recognition challenge[2] we invited participants to submit results of face detection and face recognition algorithms. Given a set of images in the training set, containing 23,349 labeled faces of 1085 known and a number of unknown persons, participants were to detect all faces in the test set, and to assign each detected face an identity of the gallery, or an *unknown* label when the algorithm decided that the person has not been labeled as known.

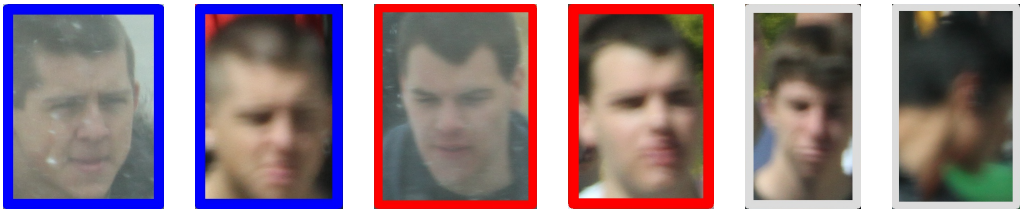
2. Dataset

To run the challenge, the UnConstrained College Students (UCCS) dataset was developed as a significantly extended version of the dataset presented by Sapkota and Boult [30]. It contains high-resolution images captured from an 18 megapixel camera at the University of Colorado Colorado Springs, aimed at capturing people walking on a

[1] https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt

[2] http://vast.uccs.edu/Opensetface

(a) Example Images



(b) Example Faces

Figure 1: EXAMPLES OF THE UCCS DATASET. Two exemplary images of the UCCS dataset including hand-annotated bounding boxes and identity labels are shown in (a). In (b) the cropped faces of the two images are displayed. Faces with the same color mark the same identity, while gray boxes mark unknown identities.

sidewalk from a long range of 100–150 meters, at one frame per second. The dataset collection was spread across 20 different days, between February 2012 and September 2013 providing a variety of images in different weather conditions such as sunny or snowy days. There are frequent occlusions due to tree branches or poles as well as sunglasses, winter caps, or fur jackets that make both detection and recognition a challenging problem. Since the captured subjects are unaware of the dataset collection and casually focus on random activities such as glancing at a mobile phone or conversing with peers while walking, there is a wide variety of face poses along with some cases of motion blur, and many cases where faces are not in the focus of the camera.

The dataset consists of more than 70000 hand-cropped face regions, which are generally larger than the actual face. An identity is manually assigned to many of the faces, where 20 % of these identities appear in two or more days. Due to the manual nature of the labeling process, for approximately 50 % of the face regions no identity could be assigned. Two example images including their manually cropped and labeled faces are shown in Fig. 1.

2.1. Protocol

We split the UCCS database into training, validation and test sets. For each image in the training and validation sets we provide a list of bounding boxes with their corresponding identity labels, including the label -1 for unknown subjects. In the test set we only supply a list of images, in which the participants need to detect the faces (face detection challenge) and provide an identity label including a similarity score to each bounding box (face recognition challenge).

To be able to evaluate the difference of recognizing unknown identities that have been seen during training (so-called *known unknowns*) and subjects that have never been seen (*unknown unknowns*) [12], we artificially *mask* several of the known identities. Some of these masked identities are present in the training, validation and test set, while some other masked identities are excluded from the training set. More details about the distribution of images and identity labels in our evaluation protocol can be found in Tab. 1.

3. Challenge Participants

Participants were invited to submit a short description of their algorithms. They are listed in the order of submission and marked with their according institution (^{a} – ^{f}, cf. list of authors on first page).

3.1. Face Detection

Baseline: The Baseline face detector uses the out-of-the-box face detector of Bob [2], which relies on boosted Local Binary Pattern (LBP) features [3] extracted from gray level images, and is trained on a combination of publicly available close-to-frontal face datasets. The implementation can

	Training		Validation		Test	
	Subjects	Faces	Subjects	Faces	Subjects	Faces
Known	1085	11012	990 (1002)	3004 (3359)	921 (954)	12636 (15312)
Unknown	?	12156	?	6688 (7575)	?	17774 (21659)
Masked in Training	117	181	102	228	116	1277
Masked not in Training	0	0	461	1189	526	4466
Total	1202	23349	1553 (1565)	11109 (12351)	1563 (1596)	36153 (40038)

Table 1: DISTRIBUTION OF SUBJECTS AND LABELED FACES. These numbers of subjects and faces are present in the evaluation protocol. The numbers in parentheses display the updated protocol (cf. Sec. 5.1). Known subjects were labeled with their (positive) ID, while unknown subjects are labeled as -1. For the masked identities, the participants were given the label -1, they could not differentiate between them and the unknown identities.

be downloaded from the Python Package Index.[3]

TinyFaces^b: The TinyFaces face detector [16] consists of a set of scale-specific mini-detectors, each of which is tuned for a predefined object size and sits on top of a fully convolutional [22] ResNet101 [13]. Each mini-detector is implemented as a convolutional filter, which takes convolutional features extracted from multiple layers as input and outputs a spatial heat map that represents detection confidence at every location. In addition, four filters have been tied to each mini-detector for bounding box regression [27]. The TinyFaces detector is trained on the training set of WIDER FACE [31]. During training multi-resolution sampling, balanced sampling [27], and hard negative mining are applied. During testing the detector works on an image pyramid, while only running mini-detectors tuned for small object size on the interpolated level. The TinyFaces algorithm was run by the LqfNet^c team.

UCCS^a: The UCCS face detector is based on the Baseline face detector from Bob [2]. Additionally to the LBP features extracted from gray-level images, color information is added in terms of converting the image to HSV color space and extracting quantized hue (H) and saturation (S) values. A face detection cascade is trained on a combination of LBP and color values, using the training set images of the MOBIO [23], SCface [10], and CelebA [21] datasets, as well as the training images of the UCCS dataset.

MTCNN^d: Faces and facial landmarks are detected by MTCNN [33]. Besides increasing the minimally detectable face bounding box to 75 pixels, the publicly available face detector[4] was used unalteredly. The MCTNN face detector was run by the CVSSP^d team.

WVUCVL^e: The WVUCVL face detection algorithm is based on detected joints on the face, inspired by CNN-based human 2D body-pose estimation methods [17, 7]. First, coordinates of the 18 main joints of the human body (e.g., shoulder center, waist, and nose) are extracted, and multi-pose estimation is applied. Based on the five joints of the face (nose, both eyes, and both ears), frontal or side face detection is applied to the boundary of these joints, and a confidence threshold is employed. To decrease the false accept rate, thresholds are set for checking the size of the bounding box of each face. Finally, a skin color detector was trained on parts of the UCCS training set, and the distance between the distribution of the skin color of the bounding box from the distribution of the training set is thresholded.

Darknet_UOW^f: The Darknet_UOW Convolutional Neural Network (CNN) model closely follows one of the publicly available[5] architectures described in [26], while adding a few modifications to accommodate for differences between the Visual Object Classes Challenge 2012 (VOC2012) [9] dataset used in [26] and the UCCS dataset. The Darknet_UOW architecture consists of 22 convolutional layers followed by 2 fully connected layers, and the input size of the network has dimensions of 416×416. Each image is divided into a 5×5 grid. For each grid cell, 5 bounding boxes are predicted. For each bounding box, the center and the dimensions are extracted, as well as a confidence represented with Intersection Over Union (IOU) between the predicted bounding box and a ground truth.

3.2. Face Recognition

Baseline: For the Baseline face recognition algorithm,[3] first the faces of the training set were re-detected, and face images of 64×80 pixels were cropped. Histogram sequences of uniform LBP patterns [1] with 16×16 pixel block sizes are extracted. A Linear Discriminant Analysis (LDA) was performed on PCA-projected features [34], using all features of unknown identities (-1) in one class, and each known identity in a separate class. For enrollment of a subject (including -1), an average of the training set features is computed. At test time, LBPHS features of detected faces are projected into the PCA+LDA subspace, and cosine similarities to gallery templates are computed.

LqfNet^c: A 32×32 pixel low-resolution CNN [15] is used to project each detected and downscaled face image to a discriminative 128-dimensional face descriptor. The max-margin based network training incorporates data aug-

[3] http://pypi.python.org/pypi/challenge.uccs

[4] http://github.com/kpzhang93/MTCNN_face_detection_alignment

[5] http://pjreddie.com/darknet

mentation strategies such as blurring or adding noise to adjust the high quality training data to the low-quality domain [14]. About 9M face images from different public and private datasets serve as training data for the Low-quality face Network (LqfNet), while no training on the challenge data is performed. Similar to the Baseline, for identification an LDA is learned on the gallery descriptors. Because the LqfNet is not specifically designed to handle misdetections, the descriptor distance d is weighted by the detection confidence c to shape the final recognition score $s = c/d$.

UCCS[a]: The UCCS contribution relies on features from the publicly available[6] VGG Face descriptor network [24], which are extracted of 224×224 pixel cropped images. The enrollment is based on the Extreme Value Machine (EVM) [28], which is particularly designed for open-set recognition. Distributions of cosine distances between deep features of different identities are modeled using concepts of Extreme Value Theory (EVT), and a *probability of inclusion* is computed for each enrollment feature. Set-cover [28] merges several features of one identity into a single model, including a model for the unknown identities (-1). We optimized EVM parameters [12] on the validation set. For a probe bounding box, VGG Face descriptors are computed, and the cosine similarities between probe feature and model features are multiplied with the probability of inclusion of the corresponding EVM model.

CVSSP[d]: Features are extracted by two 29 layer CNNs. The first network is trained on combined CASIA-Webface [32] and UMD [4] face datasets, while the other is trained on CASIA-Webface, UMD and PaSC [5] datasets. Two feature vectors are extracted from each face and its mirror image, and merged by element-wise summation. The template for each enrolled subject is the average of the face features extracted from the gallery images. For the unknown subjects (-1), the face features are used as cohort samples for test-normalization. During testing, the face features are extracted and the cosine similarity scores between the templates and cohort samples are computed. For the class of unknown subjects, the similarity score is the negative of the minimum of all the templates scores. In total, there are 1086 scores for each probe face and those scores are normalized by test-normalization. The final score is the average of the two CNN models.

4. Evaluation

The face detection evaluation is performed on the test set where the participants provided detected face bounding boxes including confidence scores. For face recognition, participants turned in up to ten identity predictions for each bounding box along with a similarity score for each prediction. The evaluation scripts[3] for the validation set were given to the participants. For all our evaluations, colors across plots correspond to identical participants.

[6] http://www.robots.ox.ac.uk/~vgg/software/vgg_face

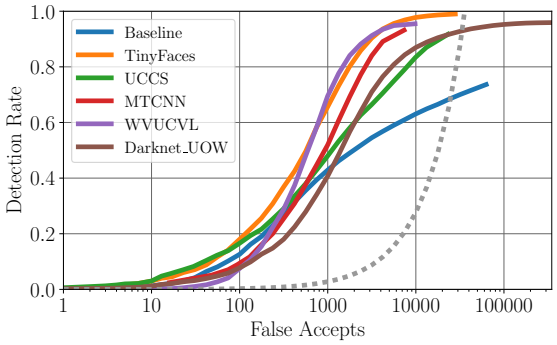


Figure 2: FACE DETECTION EVALUATION. A Free Response Operating Characteristics (FROC) curve is shown for the test set. The horizontal axis includes the number of false accepts (misdetections), while the vertical axis outlines the relative number of detected faces. The dotted gray line indicates equal numbers of correct detections and false accepts.

4.1. Face Detection

To evaluate the correctness of each bounding box, we use a modification of the Jaccard index, where the original Jaccard index is known as the Intersection Over Union (IOU) of the ground truth and the detected bounding box. As the ground truth is up to four times larger than the face (cf. Fig. 1(b)), we modify the union term to not penalize detections smaller than the ground truth:

$$J(G, D) = \frac{|G \cap D|}{\max\left\{\frac{|G|}{4}, |G \cap D|\right\} + |D| - |G \cap D|} \approx \frac{|G \cap D|}{|G \cup D|} \quad (1)$$

where D is the area of detected bounding box and G is the area of ground-truth bounding box. Hence, when the detected bounding box D covers at least a fourth of the ground-truth bounding box G and is entirely contained in G, modified Jaccard index $J = 1$ is achieved. In our evaluation, we accept all bounding boxes with a modified Jaccard index $J \geq 0.5$.

For the face detection evaluation, bounding boxes along with their confidence scores are used in a Free Response Operator Characteristic (FROC) curve [8]. Particularly, we split the confidence scores c into positives C^+, i.e., where the detected bounding box overlaps with a ground truth according to (1), and the negatives C^- where $J(G, D) < 0.5$ for each ground-truth bounding box G. For a given number of false accepts FA, we compute a confidence threshold θ:

$$\theta = \arg\max_{\theta'} \left|\{c \mid c \in C^- \wedge c \geq \theta'\}\right| < FA\,. \quad (2)$$

Using this threshold, the detection rate DR is computed as the relative number of detected faces where the detection

FA/FI	Baseline	TinyFaces	UCCS	MTCNN	WVUCVL	Darknet_UOW	Baseline	LqfNet	UCCS	CVSSP
10	460	1069	**1093**	0	43	405	13	11	**84**	13
100	4560	**6514**	5962	3211	2720	2807	108	98	**374**	84
1000	15506	23733	17350	18772	**25129**	14734	677	750	**3178**	648
10000	22802	**35349**	30139	33691	34519	31443	2696	5349	**6858**	6694
100000	26625	**35789**	33152	33691	34519	34434	3473	8920	6858	**11274**

Table 2: FACE DETECTION AND RECOGNITION RESULTS. The number of detected (left) and correctly identified (right) faces of the test set are presented for certain numbers of false accepts or false identifications, respectively. For algorithms with fewer false accepts/identifications, the total number of detected/identified faces is given. The best results are highlighted in color.

confidence is above threshold:

$$DR(\theta) = \frac{\left|\{c \mid c \in C^+ \wedge c \geq \theta\}\right|}{M}, \qquad (3)$$

where M is the total number of labeled faces given in Tab. 1. Finally, the FROC curve plots the DR over the number of false accepts, for given values of FA.

Fig. 2 presents the results of the participants on test set, while Tab. 2 contains more detailed results. Despite the difficulty of the dataset, all face detectors (besides the Baseline) detected at least 33000 of the 36153 labeled test set faces. Honestly, we (the challenge evaluators) were positively surprised by this result. However, these high results can only be achieved with a relative high number of false accepts. Still, the best performing face detectors (TinyFaces and WVUCVL) detected more than 23000 faces with 1000 false accepts, which – given the difficulty of many of the faces – is a very good result. On the other hand, assuming 5 faces in each section of a 5×5 grid cell in the Darknet_UOW algorithm leads to a large amount of false accepts, and might have missed some faces, i.e., when more than 5 faces were present in a certain grid cell.

4.2. Face Recognition

To participate in the face recognition challenge, participants were given only the raw images, i.e, without any labels. In such an open-set scenario, a face recognition algorithm has three goals. First – similarly to closed-set identification – if the probe face is of a known identity, the corresponding gallery template of that identity must have the highest similarity across all gallery templates. Second, if the probe face is of an unknown identity, the similarities to *all* gallery templates need to be small, or the probe should be labeled unknown. Finally, when the face detector has a misdetection, that region should be handled as unknown.

For our face recognition challenge we evaluate the participants using the Detection and Identification Rate (DIR) curve [25] on rank 1. For each probe image, we split the similarity scores s of the provided bounding boxes into two groups: positives S^+ and negatives S^-. The positive group S^+ contains similarity scores of correct identifications. For each ground-truth bounding box of a known identity, the detected bounding box with the highest overlap is considered,

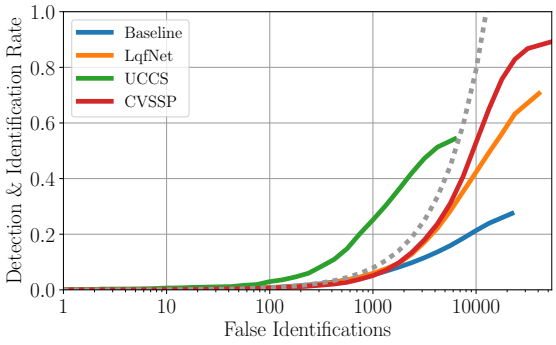


Figure 3: FACE RECOGNITION EVALUATION. A Detection and Identification Rate (DIR) curve at rank 1 is shown for the test set. The horizontal axis includes the number of false identifications, while the vertical axis outlines the relative number of correctly identified faces. The dotted gray line indicates equal numbers of correct and false identifications.

but only if the modified Jaccard index exceeds the overlap threshold $J \geq 0.5$ defined in Sec. 4.1. If the assigned subject label with the highest similarity score s of that bounding box corresponds to the correct identity, s is added to S^+.

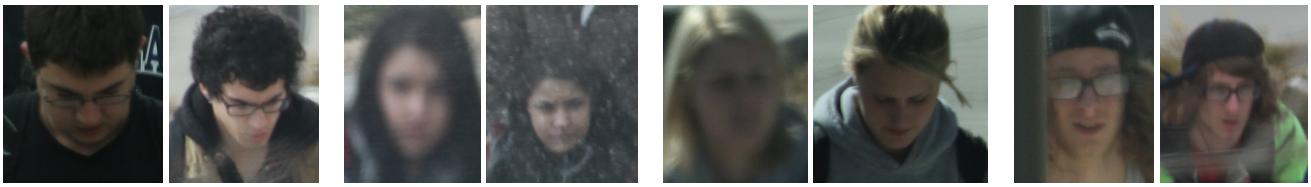
The negative group S^- contains similarity scores of false identifications. It is composed of the unknown face images and the false accepts, which are labeled as a known identity. For each ground-truth bounding box of an unknown identity, the detected bounding box with the highest overlap was considered. If the assigned subject label with the highest similarity score s of that bounding is *not* -1, s is added to S^-. For false accepts, i.e., where the modified Jaccard index to every ground-truth bounding box is lower than $J < 0.5$, and where the highest similarity score s of that bounding box is not labeled -1, s is appended to S^-.

A decision threshold ϑ on the similarity scores can be computed for a given number of false identifications FI:

$$\vartheta(FI) = \arg\max_{\vartheta'} \left|\{s \mid s \in S^- \wedge s \geq \vartheta'\}\right| < FI\,. \qquad (4)$$

Using this decision threshold, the relative number of correctly detected and identified persons is computed as:

$$DIR(\vartheta) = \frac{\left|\{s \mid s \in S^+ \wedge s \geq \vartheta\}\right|}{N}, \qquad (5)$$



(a) (b) (c) (d)

Figure 4: AUTOMATICALLY ASSIGNED IDENTITIES. Some examples of automatically assigned identities are presented at the left of each pair, together with a corresponding gallery face of the newly assigned identity. New faces are assigned when all three competitors agree on the same identity. The identity in (a) is assigned wrongly, while the remaining identities are correct. The gallery and probe images shown in (d) were taken at different days.

where N is the number of known faces, cf. Tab. 1. Finally, the DIR curve plots the detection and identification rate over the number of false identifications, for selected values of FI. In our DIR plots, we do not normalize the false identifications by the total number of unknown probe faces as done by Phillips *et al.* [25]. As our group of false identifications $\mathcal{I}^-$ includes false accepts, which are different for each face detector, normalization would favor participants with a high number of false accepts.

Fig. 3 and Tab. 2 present the results of the participants on the test set. Given the difficulty of the dataset, from a closed-set perspective the results are impressive: almost 90 % of the faces were correctly identified at rank 1 by CVSSP. From an open-set perspective, however, this comes at the price of more than 50000 false identifications. LqfNet did not reach such a high identification rate, but still produced around 40000 false identifications. Hence, for both algorithms up to 5 times more false identifications are raised than people are identified. Eventhough the number of probe faces containing unknown identities is higher than the number of probes of known identities (cf. Tab. 1), the number of false identifications is far too high to be usable in a real scenario. On the other hand, the UCCS algorithm, which is the only algorithm that can identify people with lower numbers of false identifications, reaches only an identification accuracy of around 50 %.

5. Discussion

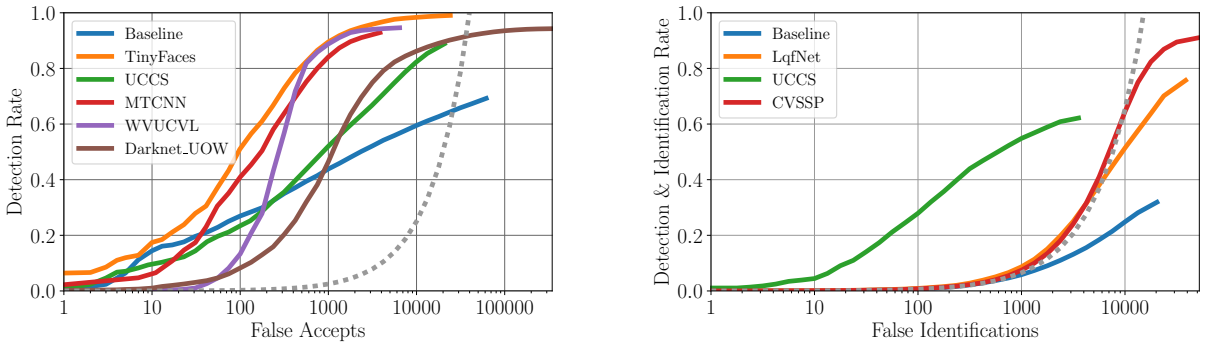
5.1. Database Clean-up

The UCCS dataset is manually labeled, both the face bounding boxes and the identities. We are aware that there are some errors in the labels, including non-marked faces in the face detection challenge, as well as faces that are labeled as -1, but which are actually one of the known gallery identities. To clean up the dataset and add face bounding boxes as well as identity labels, we opted for an automatic process – given the short time to write up this paper. The results of this automatic process are given in parentheses in Tab. 1.

To automatically mark missing faces, we use the face detectors of the participants, excluding the Baseline detector. We select those bounding boxes that are detected by the majority of face detectors with high confidence, i.e., where detections of three algorithms overlap with IOU ≥ 0.25. For each of the detectors, we computed a separate threshold θ at 2500 false accepts in the validation set. To generate a new bounding box, we merged the overlapping detections, weighted with their respectively normalized detection confidence, and up-scaled them by a factor of 1.2. In this way, we added 1242 face bounding boxes in the validation set, and 3885 faces in the test set, which we labeled as unknown. We have manually checked around 100 images with automatically added bounding boxes, and all of them contain valid faces. Still, we have found that some of the faces are not marked by this automatic process. However, for lower confidence thresholds, we found that some overlapping detections do not contain faces.

After adding these new unknown faces into the dataset, we automatically tested all faces, for which the identity was unknown. If all three face recognition algorithms agree on the same known identity on rank 1, we assign this identity to that face. Using this technique, 355 faces in the validation set are assigned to known identities, while in the test set it amounts to 2676 faces. Manually checking around 100 newly assigned faces in the validation set, we found exactly one face, where the label is incorrectly assigned. This face is shown in Fig. 4(a), including one of the gallery faces of the assigned identity. On the other hand, all the other inspected faces are correctly labeled, including images from different days, see Fig. 4(d) for an example. Note that we could not update the masked identities, cf. Tab. 1.

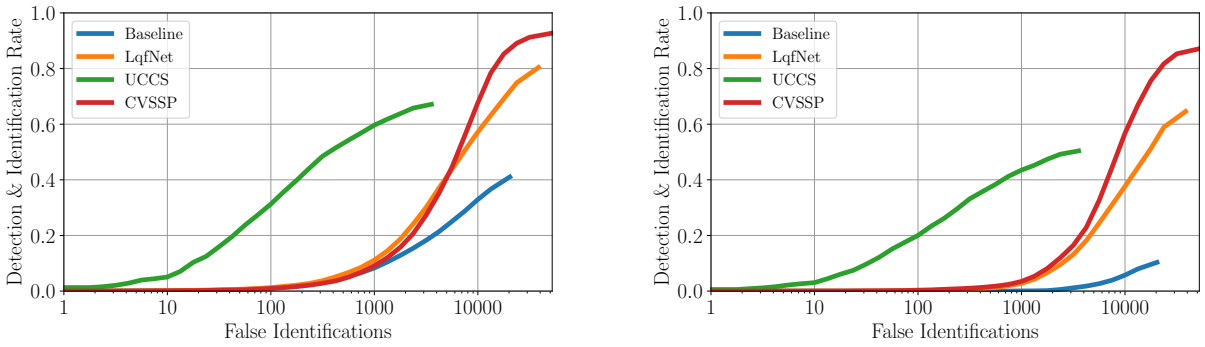
When evaluating the participants' detection and recognition algorithms on the updated labels of the test set, we can see that the results generally improve. In the FROC curve in Fig. 5(a), out of the 40038 faces that are labeled in the updated test set, two algorithms can detect around 90 % of them at 1000 false accepts. Interestingly, in comparison to Fig. 2, MTCNN is almost able to catch up with TinyFaces and WVUCVL. Finally, TinyFaces detects almost all faces, but on the cost of more than 30000 false accepts.



(a) Face Detection

(b) Face Recognition

Figure 5: UPDATED EVALUATION. FROC and DIR curves of all participants are displayed on the test set for the automatically updated ground-truth labels. The dotted gray line indicates equal numbers of correct detections/identifications and false accepts/identifications.



(a) Same Day

(b) Different Day

Figure 6: SAME VS. DIFFERENT DAY. DIR curves are comparing templates and probes taken on (a) the same and (b) different days. The horizontal axis includes the number of false identifications, while the vertical axis outlines the rate of correctly identified faces.

The DIR curve in Fig. 5(b) does not show any differences to Fig. 3 with respect to the order of the participants, as all participants need to agree for a face to be re-labeled. However, we can see an improvement of all algorithms, and for 1000 false identifications now all algorithms are over the gray dotted line, i.e., at that point all algorithms correctly identify more faces than they label by mistake.

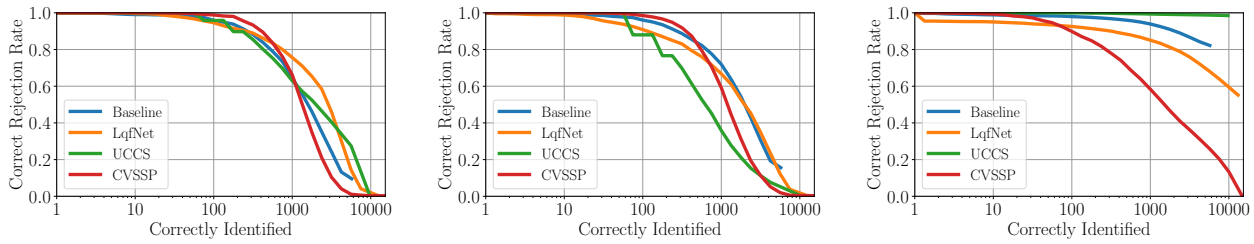
5.2. Analysis of Time Differences

Though the UCCS dataset was collected over more than two years, many people appear only in a single day. The training and validation sets are built such that for most of the known identities all faces stem from a single day. In the test set, we have put some images of known subjects, which are taken at a different day than present in the training set. The updated protocol of the test set contains 20647 faces of 932 identities taken at the same day as the training set faces of the corresponding identities, and 14432 faces of 209 identities with different days.

In Fig. 6 we show the difference of the participants' results between the two sets of probe faces in the updated test set. As we could not split the unknown faces into same or different days, the same false identifications are used in both plots. It comes as no surprise that the different day faces are harder to be identified, the rates in Fig. 6(b) are generally lower than in Fig. 6(a). However, the behavior of the algorithms is different. While the drop for CVSSP is moderate, both LqfNet and UCCS decrease identification rates up to 20 %, and the Baseline is practically useless when images are taken at different days.

5.3. Analysis of False Identifications

In the evaluations above, false identifications are computed jointly from unknown faces and false accepts (misdetections). To evaluate the impact of each of these on the performance, we split the false identifications. Here, we only use the *masked* faces, i.e., which have been given as -1 to the participants, but where we know the identity la-



(a) Masked in Training (b) Masked not in Training (c) False Accepts

Figure 7: EVALUATION OF FALSE IDENTIFICATIONS. In (a) and (b) the percentage of correctly rejected masked identities are plotted over the number of correctly identified faces. In (c) the relative number of correctly rejected false accepts (misdetections) is plotted over the number of correctly identified faces, normalized by the total number of false accepts per detection algorithm.

bels. The masked faces are further split up into masked identities that are in the training set, and masked identities that are not, cf. Tab. 1. The evaluation is performed on the automatically cleaned dataset.

To have a better comparable evaluation, we plot correctly rejected masked faces or false accepts over the correctly identified known identities. Hence, the similarity score threshold ϑ is now computed over $\mathcal{S}^+$ (cf. Sec. 4.2), while the correct rejection rate is computed as:

$$CRR(\vartheta) = \frac{\left|\{s \mid s \in \mathcal{S}^= \wedge s < \vartheta\}\right|}{\left|\{\mathcal{S}^=\}\right|} \tag{6}$$

where $\mathcal{S}^= \subset S^-$ are the corresponding false identifications.

In Fig. 7(a) we plot the number of correctly rejected masked images where the identities are included in the training set, while Fig. 7(b) contains the masked images where subjects are not part of the training set. As algorithms that model unknown faces as a separate class during training, the LqfNet and UCCS algorithms decrease rejection capabilities for unknown identities unseen during training (Fig. 7(b)). On the other hand, the CVSSP algorithm did not make use of the training set and, thus, its performance is stable with respect to the masked identities.

Finally, Fig. 7(c) shows how the algorithms deal with their respective false accepts. There, the UCCS algorithm, which particularly models rejection probabilities, is able to reject almost all false accepts, even with low threshold ϑ (high number of correct identifications), while LqfNet rejects half of the false accepts to be unknown, and CVSSP assigns a known identity label to each of its false accepts.

6. Conclusion

We have evaluated the participants' results of the unconstrained face detection and the open-set face recognition challenge. We were surprised by the quality of the face detectors, and the closed-set recognition capabilities of the algorithms on our very difficult dataset. However, open-set face recognition, i.e., when face recognition algorithms are confronted with unknown faces and misdetections is far from being solved. Either the algorithms achieved their good identification accuracies only at high false identification rates, or they identified only up to 60 % of the faces.

For this paper, we automatically updated our ground-truth labels by majority voting of the participants' algorithms. With this, we surely have missed some of the faces, and some identity labels are definitely wrong. We will use the participants' results to start a semi-manual re-labeling of the data, i.e., we propose overlapping bounding boxes to a human observer who decides whether a face is seen, or whether two face images show the same identity. The training and validation sets will be made public after this process is finished, while the test set will be kept secret and used in further challenges.

For the present competition, we provided a biased evaluation protocol, i.e., the training set is identical to the enrollment set. As we have seen, already with this biased protocol open-set face recognition is difficult. More unbiased evaluations would split off several identities into a training set, and enrollment and probing would be performed on a different set of identities. We will investigate on such an unbiased protocol in future work.

Acknowledgment

This research is based upon work funded in part by NSF IIS-1320956 and in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face recognition with local binary patterns. In *European Conference on Computer Vision (ECCV)*. Springer, 2004.

[2] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel. Bob: a free signal processing and machine learning toolbox for researchers. In *ACM Conference on Multimedia Systems (ACMMM)*, 2012.

[3] C. Atanasoaei. *Multivariate Boosting with Look-up Tables for Face Processing*. PhD thesis, École Polytechnique Fédérale de Lausanne (EPFL), 2012.

[4] A. Bansal, A. Nanduri, R. Ranjan, C. D. Castillo, and R. Chellappa. UMDFaces: An annotated face dataset for training deep networks. In *Arxiv Preprint arXiv:1611.01484*, November 2016.

[5] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2013.

[6] J. R. Beveridge, H. Zhang, B. A. Draper, P. J. Flynn, Z. Feng, P. Huber, J. Kittler, Z. Huang, S. Li, Y. Li, V. Štruc, J. Križaj, et al. Report on the FG 2015 video person recognition evaluation. In *International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2015.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Conference of Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[8] D. Chakraborty. Statistical power in observer-performance studies: Comparison of the receiver operating characteristic and free-response methods in tasks involving localization. *Academic Radiology*, 9(2), 2002.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision (IJCV)*, 88(2), 2010.

[10] M. Grgic, K. Delac, and S. Grgic. SCface — surveillance cameras face database. *Multimedia Tools and Applications*, 51(3), 2011.

[11] M. Günther, A. Costa-Pazo, C. Ding, E. Boutellaa, G. Chiachia, H. Zhang, M. de Assis Angeloni, V. Struc, E. Khoury, E. Vazquez-Fernandez, D. Tao, M. Bengherabi, D. Cox, S. Kiranyaz, T. de Freitas Pereira, J. Zganec-Gros, E. Argones-Rúa, N. Pinto, M. Gabbouj, F. Simões, S. Dobrisek, D. González-Jiménez, A. Rocha, M. Uliani Neto, N. Pavesic, A. Falcão, R. Violato, and S. Marcel. The 2013 face recognition evaluation in mobile environment. In *International Conference on Biometrics (ICB)*. IAPR, IEEE, 2013.

[12] M. Günther, S. Cruz, E. M. Rudd, and T. E. Boult. Toward open-set face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, 2017.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[14] C. Herrmann, D. Willersinn, and J. Beyerer. Low-quality video face recognition with deep networks and polygonal chain distance. In *Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2016.

[15] C. Herrmann, D. Willersinn, and J. Beyerer. Low-resolution convolutional neural networks for video face recognition. In *Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2016.

[16] P. Hu and D. Ramanan. Finding tiny faces. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[17] U. Iqbal and J. Gall. Multi-person pose estimation with local joint-to-person associations. *European Conference on Computer Vision (ECCV)*, 2016.

[18] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The MegaFace benchmark: 1 million faces for recognition at scale. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[19] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[20] J. C. Klontz and A. K. Jain. A case study on unconstrained facial recognition using the boston marathon bombings suspects. Technical Report MSU-CSE-13-4, Department of Computer Science, Michigan State University, 2013.

[21] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*. CVF, IEEE, 2015.

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[23] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J.-F. Bonastre, P. Tresadern, and T. Cootes. Bi-modal person recognition on a mobile phone: using mobile phone data. In *International Conference on Multimedia & Expo (ICME) Workshop*. IEEE, 2012.

[24] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference (BMVC)*, 2015.

[25] P. J. Phillips, P. Grother, and R. Micheals. *Handbook of Face Recognition*, chapter Evaluation Methods in Face Recognition. Springer, 2nd edition, 2011.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Conference of Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[28] E. M. Rudd, L. P. Jain, W. J. Scheirer, and T. E. Boult. The extreme value machine. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. To appear.

[29] J. Sanchez del Rio, D. Moctezuma, C. Conde, I. Martin de Diego, and E. Cabello. Automated border control e-gates and facial recognition systems. *Computers & Security*, 62, 2016.

[30] A. Sapkota and T. E. Boult. Large scale unconstrained open set face database. In *International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2013.

[31] S. Yang, P. Luo, C.-C. Loy, and X. Tang. WIDER FACE: A face detection benchmark. In *International Conference on Computer Vision (ICCV)*. IEEE, 2016.

[32] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *ArXiv preprint arXiv:1411.7923*, 2014.

[33] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *Signal Processing Letters*, 23(10), 2016.

[34] W. Zhao, A. Krishnaswamy, R. Chellappa, D. L. Swets, and J. Weng. Discriminant analysis of principal components for face recognition. In *Face Recognition: From Theory to Applications*. Springer, 1998.