“© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”

Report on the BTAS 2016 Video Person Recognition Evaluation

Walter J. Scheirer[1] Patrick J. Flynn[1] Changxing Ding[2] Guodong Guo[3] Vitomir Štruc[4]
Mohamad Al Jazaery[3] Klemen Grm[4] Simon Dobrisek[4] Dacheng Tao[2] Yu Zhu[3]
Joel Brogan[1] Sandipan Banerjee[1] Aparna Bharati[1] Brandon RichardWebster[1]

[1]University of Notre Dame
Notre Dame, IN, USA
{walter.scheirer, flynn}@nd.edu

[2]University of Technology, Sydney
Sydney, Australia

[3]West Virginia University
Morgantown, WV, USA

[4]University of Ljubljana
Ljubljana, Slovenia

Abstract

This report presents results from the Video Person Recognition Evaluation held in conjunction with the 8th IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS). Two experiments required algorithms to recognize people in videos from the Point-and-Shoot Face Recognition Challenge Problem (PaSC). The first consisted of videos from a tripod mounted high quality video camera. The second contained videos acquired from 5 different handheld video cameras. There were 1,401 videos in each experiment of 265 subjects. The subjects, the scenes, and the actions carried out by the people are the same in both experiments. An additional experiment required algorithms to recognize people in videos from the Video Database of Moving Faces and People (VDMFP). There were 958 videos in this experiment of 297 subjects. Four groups from around the world participated in the evaluation. The top verification rate for PaSC from this evaluation is 0.98 *at a false accept rate of* 0.01 *— a remarkable advancement in performance from the competition held at FG 2015.*

Figure 1. Clips of two people sampled from four PaSC handheld videos: files 06599d91.mp4, 06599d451.mp4, 05450d1359.mp4 and 05450d1759.mp4.

1. Introduction

Recognizing people in videos is challenging, and to a large extent current approaches focus on finding and recognizing the faces of the people in the videos. To better capture and share how current methods perform on video face recognition, we present the results from the BTAS 2016 Video Person Recognition Evaluation. In this evaluation, four groups (three competitors and one group establishing a new baseline for the evaluation) participated by developing algorithms and contributing results on three experiments: high-quality (control) video, handheld video and video of moving faces and people. There are two innovations over previous evaluations in this series. The first measures algorithm performance on two video datasets collected at different institutions. By incorporating two qualitatively different datasets, the competition measured the ability of algorithms to generalize across datasets. The second compared human and algorithm performance on videos from two datasets.

By design, many of the complications that arise in video face recognition are amply represented in the Point-and-Shoot Challenge Face Recognition Challenge (PaSC) [3]; the BTAS 2016 Video Person Recognition Evaluation con-

sists of two experiments from the PaSC. The videos in the PaSC data set show people in motion carrying out actions; the goal is to recognize the people performing the actions, not to recognize the actions. In addition, the videos are acquired using several different grades of cameras in a variety of settings both indoors and outdoors. The result is a set of video-to-video person recognition instances ranging from relatively easy to extremely challenging. Four sample frames from the PaSC video data appear in Figure 1.

The Video Database of Moving Faces and People (VDMFP) [16] was developed for evaluating the effects of motion on human memory for faces and people, but was also found to be useful for testing recognition algorithms. The VDMFP videos were collected in two scenarios. In the first a subject walks towards the camera. In the second, the subject to be recognized is talking with another person. The camera is looking down on conversation. Importantly, extensive human performance data on the VDMFP videos is available [17]. VDMFP videos were included in the Multiple Biometric Grand Challenge (MBGC) [19].

The BTAS 2016 Video Person Recognition Evaluation builds upon the International Conference on Automatic Face and Gesture Recognition (FG) 2015 and International Joint Conference on Biometrics (IJCB) 2014 evaluations in this same series [5, 4]. In particular, the first and second experiments presented here for high-quality and handheld video recognition are identical to the experiments in the FG 2015 competition. The top verification rate at FAR=0.01 for this evaluation is a remarkable leap beyond the top performer in the prior competition, a jump from 0.58 to 0.98. These results demonstrate a major advance in algorithm design since the last evaluation, and it is now evident that the PaSC video no longer represents a significant challenge for deep learning-based approaches. However, reported results for VDMFP are not nearly as strong, raising the question of the ability of learning-based algorithms to generalize across datasets [24].

2. Related Work

The YouTube Faces dataset is a popular dataset that consists of $3,425$ videos of $1,595$ people collected from YouTube [26]. Since the videos are from YouTube, they were taken using a variety of settings and sensors. On this data set, performance is measure on a verification task and the measure of accuracy is $1 - \text{EER}$, where EER is the equal error rate. At the time this paper was written, the highest reported accuracy was 91.4% for the DeepFace algorithm [23].

The IJCB 2014 PaSC Video Face and Person Recognition Competition [5] reported the performance in a still image-to-video experiment and a handheld video experiment, the latter being the same as the handheld experiment reported here.

(a) (b)

Figure 2. Example frames from video sequences in the VDMFP. The image in (a) is from a video sequence of a subject walking towards the cameras in an atrium. The image in (b) is acquired from a video camera looking down on a conversation. In (b), the subject whose face can be seen is the subject to be recognized.

The FG 2015 Video Person Recognition Evaluation [4] reported performance for two video-to-video matching problems from the PaSC dataset. The first consisted of videos from a tripod mounted high quality video camera (control). The second contained videos acquired from 5 different handheld video cameras. Both of these experiments are included in the 2016 evaluation described here.

3. Data, Experiments and Protocol

3.1. Video Data–PaSC

The videos in the PaSC dataset were acquired in seven weeks during the Spring 2011 academic semester at the University of Notre Dame. During each week, all subjects performed the same scripted action (out of seven total actions), which changed from week to week. A handheld and control video was acquired at the same time for each subject. Thus, there is a one-to-one correspondence in terms of subject and action between handheld and control videos. Handheld videos were acquired by five cameras (one model per week) and the control videos from the same week were acquired by the same camera.

3.2. Video Data–VDMFP

The Video Database of Moving Faces and People (VDMFP) was collected at the University of Texas at Dallas, in hallways and atria with unconstrained pose and illumination, as shown in Figure 2. The dataset contains two types of video sequences: walking and conversation. In the walking sequences, the subjects walked towards the camera. In the conversation sequences, a camera is looking down on a conversation between two subjects. One subject's back is to the camera and this subject is to be ignored in the experiments. The other subject's face is facing the camera at an off angle. This is the subject to be recognized.

3.3. Experiments and Protocol

The protocol for this evaluation asked participants to deliver to the organizers three similarity matrices. These matrices contain similarity scores generated by the participants' matching algorithms. Each entry in the matrix contains a score $s(q,t)$ that is the similarity between videos q and t as generated by the participants' matching algorithm. These matrices are in a format originally developed by the National Institute of Standards and Technology, and support code to help work with these matrices and is included in the PaSC Software Support Package[1]. Participants delivered these matrices and receiver operating characteristics (ROC) curves to the organizers. The organizers worked with the participants to confirm the matrices were in the correct format and that the organizers could reproduce the ROC curves from the similarity matrices.

The three similarity matrices correspond to these three experiments in the evaluation:

1 **Control:** Compare all $1,401$ control videos to each other and generate the complete set of possible similarity scores ($1,962,801$ similarity scores).

2 **Handheld:** Compare all $1,401$ handheld videos to each other and generate the complete set of possible similarity scores.

3 **VDMFP:** Compare all 958 videos to each other and generate a similarity score matrix containing all $917,764$ possible similarity scores.

In all three experiments, all videos are compared to all videos; this maximizes the number of comparisons possible. The protocol includes the degenerate case along the diagonal of the matrix where videos are compared to themselves, which were ignored in our analysis. A video-pair is a match pair if the person in both videos is the same and a video-pair is a non-match pair if the people are different.

This evaluation followed the PaSC protocol. The PaSC protocol placed limitations on the training set and the use of cohort or gallery normalization. Algorithm training sets cannot include videos in the evaluation data set, imagery of subjects included in the PaSC dataset, or data collected at the University of Notre Dame in the Spring 2011 semester. The last restriction prevents training algorithms on environments in the PaSC dataset. The imagery for cohort or gallery normalized sets have the same restrictions.

A modest training set, 280 videos, is available with the PaSC data that follows the PaSC protocol for training sets. However, because this is data collected in other semesters at the University of Notre Dame under somewhat different circumstances, it is similar to the PaSC evaluation data in some ways and different in others. In general the organizers are assuming that many groups are training the algorithms on imagery not included the PaSC distribution.

In this evaluation, the relative performance of algorithms is compared first in terms of ROC curves and second in terms of the verification rate, also known as the true positive rate, at a false accept rate (FAR) of 0.01. The FAR=0.01 is chosen to be the best tradeoff between two opposing constraints.

[1] `http://www.pasc-eval.org/support.html`

4. Summary of Approaches

Four groups submitted results for this evaluation: one produced a baseline algorithm, and three competed in the challenge. Results were provided in the form of similarity matrices and the performance summary appears in Section 5. In addition to submitted results, groups were asked to provide brief descriptions of the approach they took. What appears below is based upon these participant provided descriptions.

4.1. University of Technology of Sydney (UTS)

UTS made multiple contributions to overcome challenges in video-based face recognition (VFR). For example, to enhance robustness of CNN features to pose variations and occlusion, a Trunk-Branch Ensemble CNN model (TBE-CNN) is proposed, which efficiently extracts complementary information from holistic face images and patches cropped around facial components. Moreover, to further promote the discriminative power of the representations learnt by TBE-CNN, a novel loss function called Mean Distance Regularized Triplet Loss (MDR-TL) is proposed. And beyond this, a tracking-based algorithm to automatically remove the irrelevant faces appearing in the background of each video clip is used. In the following, the key modules of the approach are briefly introduced.

Face Tracking Module: In real-world VFR applications, some irrelevant faces may appear in the background of a video clip. Both the irrelevant faces and the faces of interest are detected by face detection algorithms. A tracking-based algorithm is proposed to distinguish these irrelevant faces from the faces of interest. First, the detected facial bounding boxes are fed to a tracker [11]. As the face detection results are discontinuous, the tracker usually returns many tracklets from a video. The following key point is to select the tracklets that incorporate faces of interest from all tracklets. To achieve this goal, an assumption that the tracklet that includes the largest face in the video belongs to the subject of interest is made. Then, the cosine distance of this tracklet and each of the remaining tracklets is calculated, based on the CNN features returned by TBE-CNN. A safe threshold is set such that we exclude irrelevant tracklets with high confidence. Tracklets whose representations are close to that of the ground truth tracklet are saved for recognition.

TBE-CNN Module: To learn pose- and occlusion-robust face representations, previous approaches train models separately for the holistic face and image patches cropped around facial components [7, 9]. Although this strategy promotes performance, it significantly reduces the efficiency for VFR. The TBE-CNN model efficiently extracts complementary information from the holistic face image and facial components. TBE-CNN incorporates one trunk network and several branch networks. The trunk network is trained to learn face representations for holistic face images, and each branch network is trained to learn face representations for image patches cropped from one facial component. The trunk network implementation is based on GoogLeNet [22]. The GoogLeNet layers are divided into three levels: the low-level layers, middle-level layers, and high-level layers. The three layer levels successively extract features from the low to the high-level. Since low- and middle-level features represent local information, the trunk network and branch networks can share low- and middle-level layers. In comparison, high-level features represent abstract and global information; therefore, different models should have separate high-level layers. The trunk and branch networks are fused by concatenating their last convolutional layers. TBE-CNN outperforms the trunk network with only marginal increases in time and memory costs.

MDR-TL Module: Existing deep metric learning methods for face recognition include pairwise loss and triplet loss. Both methods rely on sampling effective image pairs or triplets from all possibilities. Since the optimization is based on each individual image pair or triplet, the global distribution of training samples is neglected, which has a negative impact on face recognition. To overcome this problem, the MDR-TL loss function is proposed, which regularizes the triplet loss by taking the global distribution of training samples within each batch into consideration.

Next, the training of TBE-CNN is briefly introduced. The publicly available CASIA-WebFace database [28] is adopted for CNN training. A similar training strategy to [8] is utilized. First, TBE-CNN is trained with softmax loss in a stage-wise manner. Then, TBE-CNN is fine-tuned with MDR-TL for one more epoch with a small learning rate of 0.001 to enhance its discriminative power.

The performance of the proposed approach is evaluated on the PaSC control and handheld sets. For VFR, the output of the 512-dimension bottleneck layer of TBE-CNN is adopted as the representation of each video frame. The representations of all video frames are fused by average pooling as the compact representation of the video. For video-to-video matching, the simple cosine distance metric is adopted to calculate the similarity score. Furthermore, another three models are trained with different input image resolution, depth, and structure. The similarity scores calculated by all models are fused by averaging. Results are provided under two protocols:

First, the face detection results provided by the database are directly employed for both tracking and recognition. Note that there are 60 videos where no faces were detected by the database. For all the similarity scores related to these 60 videos, a constant value of -5.0 is set. Second, a strong face detector is employed to detect the faces appearing in the 60 videos mentioned above. For the other videos, the face detection results provided by the database are still employed.

4.2. West Virginia University (WVU)

The approach by WVU has several steps. First, face detection is executed, which is done with different techniques for the two video databases, separately. For the PaSC database, a commercial face detection software was used, which works better than others based on conducted testing. In addition to the commercial software, some other face detectors were tested, such as Haar cascades from OpenCV [6] and the Constrained Local Model (CLM) face tracker [1] to get more detection results. For the VDMFP video database, a CLM based face detection/tracking approach can do better than others, thus it was adopted for the VDMFP videos. In addition to face detection, the algorithms also detected some facial landmarks, such as eye locations, which are used for face alignment. All detected faces are aligned and cropped, and resized to 256×256 pixels for further processing.

For person recognition, deep convolutional neural networks are used. A deep network with more than 20 layers is trained with a large number of face images. After training, the deep model is learned and used to extract deep features from all cropped face images in PaSC and VDMFP databases, respectively. Furthermore, another deep model called VGG [18] was utilized to extract different deep features from the faces in the two video databases as well. For the VDMFP video database, the MPI Person Body Recognition deep model has been used to extract the body features. Then those features were incorporated to calculate the final similarity matrix.

Given the extracted deep features, face matching is done based on distance measures between the deep features. Different distance measures including the cosine distance are used for distance/similarity measures. For each video, deep features from the detected frames are combined together to represent the video faces. The final similarity matrix is obtained by a weighted sum of the similarity measures from different deep features.

4.3. University of Ljubljana (Uni-Lj)

The approach of Uni-Lj used the provided PittPatt face detections to crop the facial regions from the video clips of the PaSC database. If no bounding box was provided for a

given video, the Viola-Jones face detector [25] was run and the resulting detections were used to extract facial regions from the videos. All detected faces were then rescaled to a fixed size of 224×224 pixels and subjected to a feature extraction procedure. No additional alignment step was performed on the resized images.

For feature extraction, the pre-trained VGG-Face [18] deep convolutional neural network was exploited and $4,096$-dimensional vectors were extracted from each of the detected faces. At this point of the processing pipeline each video clip was represented with a number of feature vectors. Here, the number of feature vectors varied from video to video depending on the performance of the face detector.

To compare two videos and compute a matching score for the comparison the following steps were taken:

1. The feature vectors extracted from a given video were grouped into 3 clusters using k-means clustering (with the cosine similarity measure). The similarity score between two videos was then computed by matching each of the 3 centroids of the first video against the 3 centroids of the second and computing the mean value of all valid centroid comparisons. The cosine similarity was used here for centroid matching.

2. An orthogonal subspace was computed from the feature vectors belonging to a given video. In line with the idea of the Mutual Subspace Method [27], [10] the similarity of two subspaces (representing two videos) was measured using canonical correlations. The mean value of all computed canonical correlations was taken as the final similarity score for the given comparison.

The similarity scores produced by the above procedures were ultimately combined using a weighted-sum fusion rule. For the VDMFP database the same procedure as described above was adopted. However, as no bounding boxes were provided, the Viola-Jones face detector was used with all videos of this database.

4.4. University of Notre Dame Baseline (ND)

Previous Video Person Recognition Evaluations made use of the Local Region Principal Components Analysis (LRPCA) algorithm [3], a basic subspace approach to local facial patch matching. Since the initial publication of the PaSC data set, progress in face recognition has accelerated through the use of artificial neural networks for representation learning, which far surpass basic subspace-based algorithms. For the 2016 competition, a new baseline algorithm was developed by ND that incorporated two convolutional neural networks. The motivation for the baseline algorithm's design stemmed from the need to incorporate out-of-the-box feature extraction from a popular public network (VGG-Face) that is a natural choice for this task (and indeed was used by two of the three competitors), as well as a few enhancements to account for the difficulty of both the PaSC and VDMFP datasets. Each component of the baseline algorithm is described below[2].

VGG-Face: VGG-Face is a 16-layer convolutional neural network trained on 2.6 million faces images collected from the web [18]. Eschewing decision making within the network via the softmax function, the network was instead applied as a feature extractor for the images in the challenge data sets. $4,096$ dimensional feature vectors were extracted for each face image from the fc7 layer of the network, which is located just before the final fully connected layer.

NDnet: Using the well-known AlexNet architecture [15], a feedforward convolutional neural network was trained on $175,000$ images ($7,574$ unique individuals) from the Casia-WebFace data set [28] (frontalized via the method described by Hassner et al. [13]), whose hyperparameters for training were chosen based on the HyperOpt package's random search method for hyperparameter optimization [2]. Instead of random initialization for training, all the layers of the model were fine-tuned, except the last fully connected layer (fc8), from a saved state of the same model pre-trained on ImageNet [21]. After completion of training, $4,096$ dimensional feature vectors were extracted for each face image in the challenge data sets from the fc7 layer of the network, which is located just before the final fully connected layer. These features are used as a supplement to the VGG-Face features.

Face Detection: Faces that were used as training data for NDnet were collected and annotated for facial landmarks using the method proposed by Zhu and Ramanan [29]. For the PaSC data set, the provided face coordinates were utilized. For the VDMFP data set, faces were collected and annotated for facial landmarks using the Dlib library [14]. "Bad" detections were removed in all cases by taking a sliding window over the x, y coordinates of each detection from each frame, and eliminating detections that were outside of the 1.5 sigma range (assuming "discontinuous" detections that had sporadic coordinates which were incorrect). The video frames were 2D aligned for PaSC and VDMFP after the faces were cropped out.

Video-to-Video Comparison: To prepare data for matching, features from all frames for a single video were combined via element-wise averaging. This was done for all videos in each data set. To generate matching scores, cosine similarity was computed between averaged feature vectors from VGG-Face and NDnet separately. The scores from both networks were then combined via score-level fusion. This was a weighted average of the scores from both networks, which placed most emphasis on VGG-Face (0.95 for VGG-Face, and 0.05 for NDnet).

[2]Baseline code and extracted feature vectors will be released after this work is published

5. Results

The results for the PaSC and VDMFP experiments were scored separately. UTS submitted results for the PaSC experiments and one group submitted results for two algorithms (UTS-S and UTS-P). Three groups submitted results for the VDMFP experiments.

The ROC curves for the PaSC control and handheld experiments are presented in Figure 3. The verification rates at FAR=0.01 for the five algorithms on the control and handheld videos are noted on the ROC plot. There is a wide range of performance for both experiments, with a larger range on the handheld experiment. In both experiments, the top performing algorithm had verification rates at FAR=0.01 of 0.98 to 0.97 on the control and handheld experiments respectively.

The ROC curves for the VDMFP experiment are presented in Figure 4. The subjects in the VDMFP videos follow two scripts: walking towards a camera or engaging in a conversation (see Figure 2). In the VDMFP experiment, algorithms compare videos under three conditions: walking to walking, conversation to conversation, and walking to conversation. The plot on the top shows ROCs for all three conditions. Here VR at FAR=0.01 varies from 0.02 to 0.32. The bottom plot reports results for only the walking to walking condition and VR at FAR=0.01 varies from 0.20 to 0.72. For the three algorithms in both the PaSC and VDMFP experiments, performance is comparable for the PaSC control and the VDMFP walking-to-walking experiment. This suggests that processing the conversation video is challenging. This includes detection and recognition.

6. Comparison with Human Performance

Phillips et al. [20] compared humans and algorithms on the PaSC handheld videos. Human performance was benchmarked at two levels: challenging and extremely difficult. For the extremely difficult level, 100 pairs of videos were selected so that the accuracy of the PittPatt baseline algorithm was 100% incorrect. For the challenging level, 100 video-pairs were selected so performance was roughly at chance. Human performance is reported for aggregate scoring and fusing of the ratings of 16 subjects. Accuracy of fusing human ratings is superior to aggregate scoring. Figure 5 shows ROC curves for human and algorithm performance on the two PaSC experiments. For the challenging level, algorithm Uni-Lj is comparable to aggregate human performance, UTS and WVU are superior, and UTS is comparable to fused human performance. For the extremely difficult level, algorithm WVU and Uni-Lj's performance is worse than random, which is to be expected from the video-pair selection method. UTS is comparable to aggregate human performance. Prior to this evaluation, all algorithms were worse than random for the extremely difficult level [20].

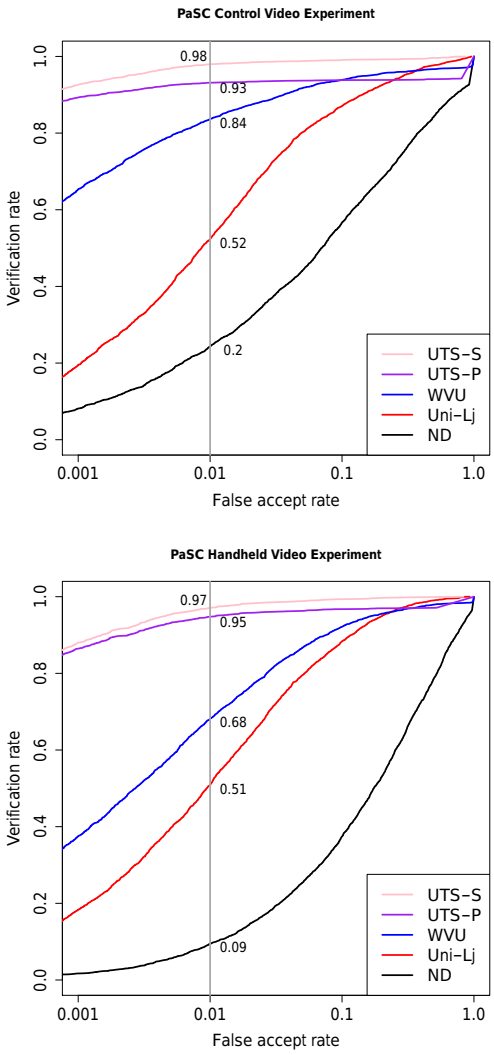

Figure 3. ROC curves for the PaSC control and handheld video experiments.

7. Analysis and Conclusion

A summary of PaSC performance from June 2013 through May 2016 is given in Figure 6. Performance is from the IJCB 2014 and Face & Gesture 2015 competitions. The initial performance of VR of 0.38 at a FAR = 0.01 was from for the PittPatt-based baseline algorithm that was reported when the PaSC dataset was release in June 2013. Performance has steadily improved since 2013. Performance is now consistently above the original PittPatt-based baseline algorithm and three algorithms from two groups have a VR at a FAR of 0.01 greater than 0.85. The inclusion of the VDMSP experiment has added a new depth to the Video Person Recognition Evaluation series of evaluations. The conversation videos in the VDMSP experiments present a new challenge for researchers, and there is still room to

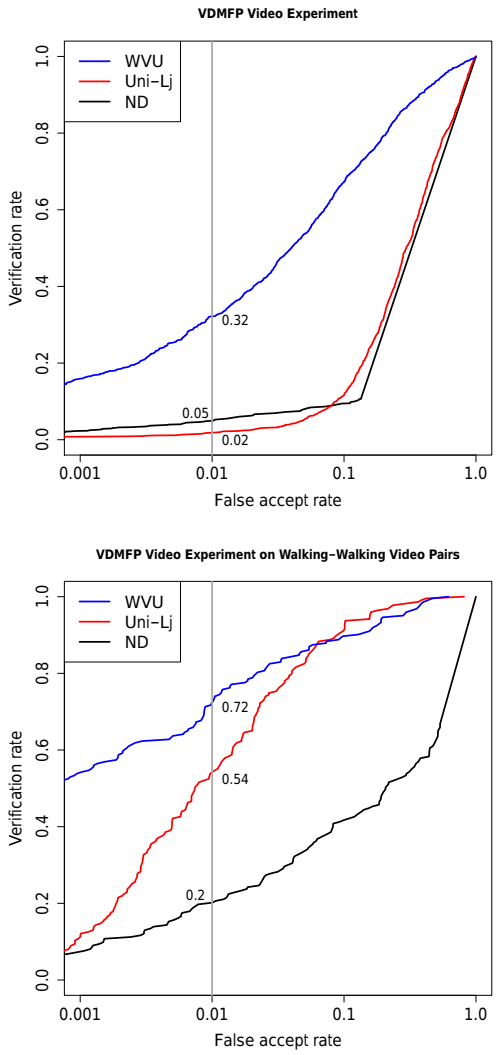

Figure 4. ROC curves for the VDMFP experiments. The ROC curves in the top plot are for the complete VDMFP videos. The bottom ROC plot is restricted to walking to walking videos.

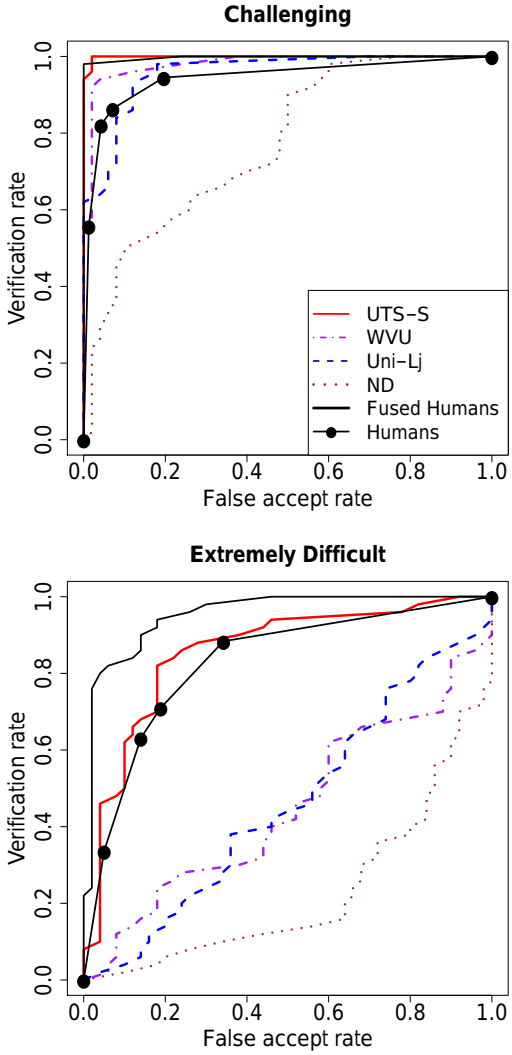

Figure 5. Human and algorithm ROCs on the challenging and extremely-difficult tasks on the PaSC videos. The legend in the challenging graph is for both graphs.

grow for even the best convolutional neural networks. The performance on the PaSC handheld and VDMSP walking to walking experiments are comparable. This suggests that advances in algorithms for the PaSC experiments transfer to similar conditions in new data sets, even though limitations with respect to generalization persist.

References

[1] T. Baltrušaitis, P. Robinson, and L.-P. Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[2] J. Bergstra, D. Yamins, and D. D. Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International Conference on Machine Learning (ICML)*, 2013.

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

[4] 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. Štruc, J. Križaj, C. Ding, D. Tao, and P. J. Phillips. Report on the FG 2015 video person recognition evaluation. In *International Conference on Automatic Face and Gesture Recognition (FG)*, 2015.

[5] J. R. Beveridge, H. Zhang, P. Flynn, Y. Lee, V. E. Liong, J. Lu, M. Angeloni, T. Pereira, H. Li, G. Hua, V. Štruc,

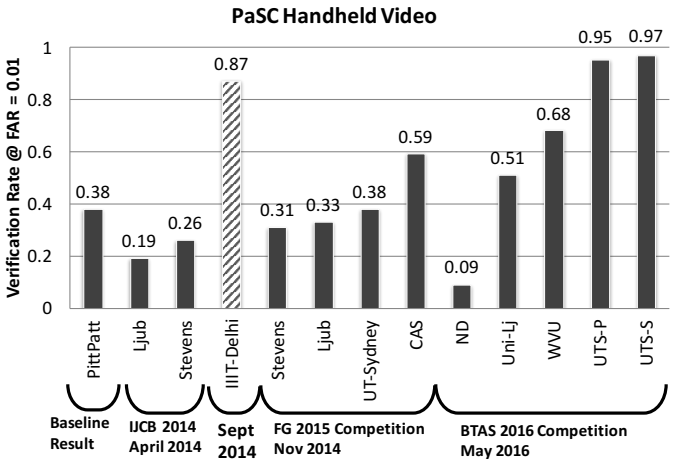

Figure 6. Summary of performance on the PaSC handheld experiment from June 2013 through May 2016. Performance reported is VR at FAR = 0.01. For performance with solid black bars, participants submitted raw scores to the PaSC organizers for scoring analysis. For the result with diagonal hashes, performance was reported in the literature [12]. For competitions, the date is when the raw scores were submitted to competition organizers. For results reported in the literature, the date is the month of publication.

J. Križaj, and P. J. Phillips. The IJCB 2014 PaSC Video Face and Person Recognition Competition. In *International Joint Conference on Biometrics (IJCB)*, September 2014.

[6] G. Bradski and A. Kaehler. *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc., 2008.

[7] C. Ding, J. Choi, D. Tao, and L. S. Davis. Multi-directional multi-level dual-cross patterns for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):518–531, 2016.

[8] C. Ding and D. Tao. Robust face recognition via multimodal deep face representation. *IEEE Transactions on Multimedia*, 17(11):2049–2058, 2015.

[9] C. Ding and D. Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on Intelligent Systems and Technology*, 7:37:1–37:42, 2016.

[10] R. Gajsek, V. Struc, and F. Mihelic. Multi-modal emotion recognition using canonical correlations and acoustic features. In *International Conference on Pattern Recognition (ICPR)*, 2010.

[11] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3d traffic scene understanding from movable platforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):1012–1025, 2014.

[12] Goswami, Gaurav and Bhardwaj, Romil and Singh, Richa and Vatsa, Mayank. MDLFace: Memorability augmented deep learning for video face recognition. In *International Joint Conference on Biometrics (IJCB)*, 2014.

[13] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[14] D. E. King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS*, 2012.

[16] A. J. O'Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, and H. Abdi. A video database of moving faces and people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):812–816, 2005.

[17] A. J. O'Toole, P. J. Phillips, S. Weimer, D. A. Roark, J. Ayyad, R. Barwick, and J. Dunlop. Recognizing people from dynamic and static faces and bodies: Dissecting identity with a fusion approach. *Vision Research*, 51(1):74–83, 2011.

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

[19] P. J. Phillips, P. J. Flynn, J. R. Beveridge, W. T. Scruggs, A. J. O'toole, D. Bolme, K. W. Bowyer, B. A. Draper, G. H. Givens, Y. M. Lui, H. Sahibzada, J. A. Scallan, III, and S. Weimer. Overview of the multiple biometrics grand challenge. In *Advances in Biometrics*, pages 705–714. Springer, 2009.

[20] 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 *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2015.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[23] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[24] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

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

[26] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 529–534, 2011.

[27] O. Yamaguchi, K. Fukui, and K.-i. Maeda. Face recognition using temporal image sequence. In *IEEE International Conference on Automatic Face and Gesture Recognition (AFGR)*, 1998.

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

[29] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.