5-2015

Report on the FG 2015 video person recognition evaluation

J.R. BEVERIDGE

H. ZHANG

B.A. DRAPER

P.J. FLYNN

Z. FENG

See next page for additional authors

Report on the FG 2015 Video Person Recognition Evaluation

J. Ross Beveridge Hao Zhang Bruce A. Draper
Colorado State University
Fort Collins, CO, USA
ross@cs.colostate.edu

Patrick J. Flynn
University of Notre Dame
Notre Dame, IN, USA

Zhenhua Feng Patrik Huber Josef Kittler
University of Surrey
United Kingdom

Zhiwu Huang[1,2] Shaoxin Li[1,2] Yan Li[1,2] Meina Kan[1] Ruiping Wang[1] Shiguang Shan[1] Xilin Chen[1]
[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences
Institute of Computing Technology, CAS, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China

Haoxiang Li Gang Hua
Stevens Institute of Technology
Hoboken, NJ, USA

Vitomir Štruc Janez Križaj
University of Ljubljana
Ljubljana, Slovenia

Changxing Ding Dacheng Tao
University of Technology, Sydney
Sydney, Australia

P. Jonathon Phillips
National Institute of Standards and Technology
Gaithersburg, MD, USA

***Abstract*— This report presents results from the Video Person Recognition Evaluation held in conjunction with the 11th IEEE International Conference on Automatic Face and Gesture Recognition. 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 1401 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. Five groups from around the world participated in the evaluation. The video handheld experiment was included in the International Joint Conference on Biometrics (IJCB) 2014 Handheld Video Face and Person Recognition Competition. The top verification rate from this evaluation is double that of the top performer in the IJCB competition. Analysis shows that the factor most effecting algorithm performance is the combination of location and action: where the video was acquired and what the person was doing.**

I. 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, here we present the results from the Face and Gesture (FG) 2015 Video Person Recognition Evaluation. In this evaluation, five groups participated by developing algorithms and contributing results on two experiments: high-quality (control) and handheld video. Video brings to face recognition a significant increase in raw data, but how useful that additional data becomes depends upon many factors, not least of which is how the people in the videos are behaving: what are they doing.

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) [2]; the FG 2015 Video Person Recognition Evaluation consists 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 FG 2015 Video Person Recognition Evaluation builds upon The International Joint Conference on Biometrics (IJCB) 2014 PaSC Video Face and Person Recognition Competition [4]. In particular, the second experiment presented here for handheld video recognition is identical to the handheld video experiment in the IJCB 2014 competition. The top verification rate at FAR=0.01 for this evaluation is double that of the top performer in the prior competition, a jump from 0.26 to 0.58. While the PaSC video still clearly remains very challenging, the results reported in this evaluation represent a major advance.

The following section briefly describes related work on video face recognition evaluation. Section III provides additional background on the PaSC [2], the two experiments included in this evaluation, and the evaluation protocol. Next, in Section IV, the approaches taken by each of the five participants are summarized. Section V presents the receiver operating characteristic (ROC) curves summarizing the performance of the five participants. Finally, in Section VI a marginal analysis of the main effects of nine

[1]CSU was funded in part by the Department of Defense through the Technical Support Working Group (TSWG).

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

covariates are reported for each participant. The covariates include properties of the faces such as size, the locations and sensors used to acquire videos, and properties of the subjects such as gender and race. The dominant factor influencing performance is the combination of locations, actions and sensors, indicating verification rates more than double when going from the most to the least challenging environments.

II. Related Work

The YouTube[2] Faces dataset is a popular data set that consists of 3425 videos of 1595 people pulled from YouTube [29]. Since the videos were pulled from YouTube, the videos were taken using a variety of settings and sensors. The measure of accuracy for this data set is $1 - \text{EER}$ for a verification task, where EER is the equal error rate. At the time this is paper was written, the highest reported performance was 91.4% for the DeepFace algorithm [24]. For the measure of accuracy we report for PaSC, the DeepFace algorithm reports a verification rate (VR) of 55% at a false accept rate (FAR) = 0.01.

The IJCB 2014 PaSC Video Face and Person Recognition Competition [4] 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. The verification rates at FAR=0.01 for the handheld video experiment for the IJCB Competition are summarize here in Table I.

TABLE I

VR @ FAR=0.01 for IJCB 2014 competition handheld video.

Pariticipant	Country	VR
Advanced Digital Science Center	Singapore	0.09
CPqD	Brasil	0.05
Stevens Institute of Technology	USA	0.26
University of Ljubljana	Slovenia	0.19

III. Data, Experiments and Protocol

A. Video Data

The videos in the PaSC were acquired in seven weeks spread out over the Spring 2011 academic semester at the University of Notre Dame. During each week, all subjects followed the same scripted action. A handheld and control video was acquired at the sametime 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 and the videos from the same week were acquired by the same camera. The locations, cameras and action combinations for the handheld video data are summarized in Table II. The control video was acquired with a Panasonic HD700 mounted on a tripod. The frame size for the control video was 1920 by 1080.

TABLE II

Handheld video location, camera and action combinations.

Sensor	**Size**	**Location**	**Action**
Flip Mino F360B	640x480	canopy	golf swing
Kodak Zi8	1280x720	canopy	bag toss
Samsung M. CAM	1280x720	office	pickup newspaper
Sanyo Xacti	1280x720	lab 1	write on easel
Sanyo Xacti	1280x720	lawn	blow bubbles
Nexus Phone	720x480	stone	ball toss
Kodak Zi8	1280x720	lab 2	pickup phone

The seven different actions were carried out according to a script - a plan. Typically, subjects began by standing at a position relatively far away and at the start of the recording started moving closer to the camera and at a diagonal relative to the camera. Then they would carry out their action, e.g. pickup a phone or toss a bean bag in a container. Finally, they would exit the scene to the side while simultaneously coming even closer to the camera. By design of the scripted actions, head size and pose changed considerably over the course of a video. These scripts meant that generally subjects were not attending to the camera, but instead looking where they were walking or concentrating on their action. In addition, videos were trimmed by hand prior to release in order to further remove portions from the start or end of the videos where subjects often stared at the person operating the camera. As a consequence of this data collection plan, while clear frontal views do arise in the videos, they are not the norm.

B. Experiments and Protocol

The protocol for this evaluation asked participants to deliver to the organizers two 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

[2]The identification of any commercial product or trade name does not imply endorsement or recommendation by NIST.

to help work with these matrices is included in the PaSC Software Support Package.[3] Participants delivered both 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 two similarity matrices correspond to the two experiments in the evaluation:

1 **Control:** Compare all 1401 control videos to each other and generate the complete set of possible similarity scores (1,962,801similarity scores).
2 **Handheld:** Compare all 1401 handheld videos to each other and generate the complete set of possible similarity scores.

In both experiments, all videos are compared to all videos; maximizing 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. For each experiment there are between 4 and 7 videos for each of the 265 people. The upper triangle of the similarity matrix contained 3128 match pairs and 977,572 non-match pairs. 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.

The FG 2015 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, 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. The imagery for cohort or gallery normalizes 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. First, in biometrics there is almost always a strong asymmetry in the cost of mistakes: generally false accepts are worse than false rejects. Thus, often FAR=0.001 is preferred [22] for mature technologies in more controlled settings. However, the video face recognition tasks in PaSC are highly challenging and the FAR=0.01 is a better choice than FAR=0.001 given the current levels of performance being seen on PaSC.

[3] http://www.cs.colostate.edu/~vision/pasc/

IV. Summary of Approaches

Five groups submitted results for this evaluation. Results were provided in the form of similarity matrices and the performance summary appears in Section V. 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.

A. Chinese Academy of Science (CAS)

In this challenge, the Chinese Academy of Science group approached the challenge using Hybrid Euclidean-and-Riemannian Metric Learning combined with deeply learned features (abbr. to HERML-DeLF), which is basically the HERML method [13] for image set classification with image features learned by a deep neural network[4]

For the feature learning part of our HERML-DeLF method, a deep convolutional neural network (DCNN) model is trained on 256 by 256 pixel face images. For a fair comparison, we normalize the face images using eye positions provided by the organizers of PaSC [2]. The DCNN model we used for feature extraction has 17 layers, i.e. 14 convolution layers, 2 fully connected layers and 1 soft-max layer. The training of the DCNN model is divided into two steps: pre-training and fine-tuning. In our work, the pre-training is conducted on "Celebrities on the Web" (CFW) database [31]. The fine-tuning is carried using two datasets. The first is the training portion of the PaSC [2]. The second is the Institute of Computing Technology, CAS-OMRON Social Solutions Co. Ltd-Xinjiang University (COX) [14] face database collected by the members of the CAS group. Finally, the output of the second fully connected layer of the trained DCNN model is used as the face feature for subsequent HERML step. Note, all the model training and feature extraction are accomplished by the Caffe deep learning framework [9] with numerous revisions specifically adapted to our face recognition task.

Using the DCNN features, the HERML method [13] is then used to compute video similarity by fusing three different set-based video representations. Specifically, for each video, the DCNN features of all video frames are first pooled respectively by sample mean, sample covariance matrix and Gaussian model, which form three types of set-based video representations. Then, by applying the kernel functions proposed in [13] for set-based representations, three kernel matrices are computed and fed separately into kernel linear discriminant analysis (KLDA) [1]. Here, instead of the original metric fusing method in [13], we exploit the KLDA to learn three projective functions respectively [28].

The resulting projective functions are then used to produce three 440 dimensional feature vectors for each video. Finally, for each pair of testing videos expressed in terms of their three KLDA feature vectors, similarity is computed as the weighted sum of three cosine similarities between the corresponding KLDA vectors. In our system, the HERML

[4] The first two CAS authors, Zhiwu Huang and Shaoxin Li, contributed equally to the development of their approach.

training is done on the training set of PaSC [2] and COX [14] face dataset.

B. University of Ljubljana (Uni-Lj)

The group from the University of Ljubljana approached the Video Person Recognition Evaluation with a recognition engine built around the MODEST framework [25]. The MODEST framework relies on probabilistic **mo**delling of **d**iverse f**e**ature **s**e**t**s. The approach is related to approaches that were part of the International Conference on Biometrics (ICB) 2013 [11] and IJCB 2014 [4] competitions.

The main idea of the MODEST framework is to represent facial images (or frames) with various texture descriptors and using the computed descriptors as input to a probabilistic modeling technique capable of deriving low-dimensional representations from the extracted texture representations.

In the first step of the Uni-Lj approach the facial region is cropped from the given video frame based on the eye coordinates provided by the competition organizers. The cropped facial region is aligned and scaled to a size of 50×50 pixels, transformed to gray-scale and then subjected to the photometric normalization technique from [27] to compensate for any potential lighting-induced artifacts in the image. The gray-scale and photometrically normalized images are then used as input for the feature extraction procedure.

During feature extraction, four different feature sets are computed/extracted from each of the two input images, i.e.:

- Gabor magnitude features, which are computed with the commonly used set of 40 Gabor filters (with 5 scales and 8 orientations) [26],
- Local binary pattern (LBP) histograms, where uniform patterns in a 8-neighborhood with a radius of 2 are used, and the local histograms are calculated from non-overlapping image blocks of size 6×6 pixels,
- Local phase quantization (LPQ) Pattern histograms, where a window size of 5 was used for the local window and the histograms were computed from non-overlapping image blocks of size 10×10 pixels, and
- Raw pixel intensities arranged into a vector that is derived from the input images by a simple concatenation of all image-rows.

As a result of this procedure, eight distinct vectors of texture descriptors are computed and subjected first to a dimensionality reduction technique and then to a modeling procedure based on a probabilistic version of linear discriminant analysis (PLDA) [18]. PLDA compresses the extracted texture information and produces low-dimensional feature vectors for each feature type.

Note that eight feature vectors are generated for each processed frame of a given video sequence. To ensure that the MODEST framework produces fixed size templates regardless of the number of frames in the video the following procedure is used. Prior to feature extraction the frames of a given video are partitioned into two groups depending on the extent of the head rotation (yaw) of the person shown in the video. Here, the first group contains frames with yaw angles below $15°$, and the second group contains frames with yaw angles greater than $15°$. Frames with negative yaw angles are mirrored prior to feature extraction to ensure that two frame-groups are sufficient to cover all rotation-dependant variability of the faces. Once the frames are grouped and the (eight) feature vectors are extracted from each frame, two pose-specific templates are constructed by averaging all feature vectors of a certain type over all frames in the given (pose) group. Hence, a template computed from the given video sequence comprises two sets of (eight) feature vectors, each feature vector having a fixed dimensionality of 200.

To produce a matching score for a given enrollment-test video pair, a template is first produced for each video. For each pose 8 partial matching scores are computed and later combined into a pose-specific similarity score using a linear combination of the partial scores. Here optimal weights for the weighted sum are learned during training using linear logistic regression (LLR). Ultimately, a single matching score is computed by averaging the two (pose-specific) scores.

C. Stevens Institute of Technology (SIT)

The Stevens Institute of Technology Group approached the video face recognition problem with the Hierarchical - Probabilistic Elastic Part (PEP) model. The Hierarchical-PEP model builds pose-invariant face representation by applying the PEP model [16], [17] hierarchically to decompose a face image into face parts at different levels of detail and thus to build pose-invariant part-based face representations. The procedure works from bottom to top in the hierarchy, stacking face part representations at each layer while reducing dimensionality in a manner that accentuates discriminative information. The Hierarchical-PEP representation of a video is a low-dimensional (100 here) vector and this size is constant across all videos; i.e. the exact number of frames in the video may vary.

The eye coordinates provided by the organizers are used to align faces and crop out 150 x 150 pixel images; i.e., aligned scaled face chips. We trained a 2-layer Hierarchical-PEP model on the Labeled Faces in the Wild (LFW) dataset [12] using the images aligned with a commercial face alignment software [30]. In the model, the first layer consists of a PEP model with 256 face part models with patch size 32 x 32 pixels. The second layer consists of PEP models with 16 face part models working on image patches of size 24 x 24 pixels. We set the dimension of the first level to 100 and the dimension of the second level to 50, hence the final face representation is of 100 dimensions.

We trained the Hierarchical-PEP model with 6,000 pairs of face images in LFW. We then construct face representations for all the 13,233 face images in LFW and train a joint Bayesian classifier [6] with their identity labels. Given two face videos, the Hierarchical-PEP model builds two 100-dimensional vectors respectively. The two vectors are then evaluated by the Joint Bayesian classifier to output a similarity score.

D. University of Surrey (Surrey)

The approach taken by the University of Surrey group tackles the PaSC video-to-video matching by combining a dynamic video frame selection method with a multi-scale local phase quantization (MLPQ) based frame-to-frame matching algorithm [5] and a simple voting strategy. By design, the frame selection method provides high-quality frames for the MLPQ-based matching algorithm to obtain a matrix of scores for each pair of videos. Then, a simple voting strategy is used to obtain the final matching score of the video pair.

A typical video contains many frames that we would like to avoid using in face recognition, e.g. frames with serious motion-blur, too few pixels on the face, or subjects looking away from the camera. Therefore, the first step in our algorithm is a frame selection that outputs the n ($n = 20$ in our results) best frames of a video, in conjunction with an associated frame score. We analyze a frame for the following quality criteria: *(1)* size of the person's face, *(2)* a sharpness score based on image edges, *(3)* a focus measure based on the Laplacian of the image, and *(4)* the orientation of the head. These individual scores are normalized and combined to form a final frame-quality score. If fewer than n frames have metadata, i.e. eye coordinates, then all available frames are used.

Given two videos with n_1 and n_2 selected frames ($n \leq 20$), we match the $n_1 \times n_2$ pairs using a MLPQ based matching algorithm. The frames are cropped to head patches and LPQ histograms are extracted from non-overlapping cells and multiple scales. Kernel discriminant analysis (KDA) is then applied before obtaining a similarity score matrix. In the last step, we sort these scores as a descending vector and use the average value of the first $\min(n_1, n_2)$ scores as the final output of the similarity of a pair of videos. The eye coordinates provided by the competition organizers are used in the frame selection and face alignment.

As a by-product of this challenge, a set of 5 and 68 facial landmarks obtained with our random cascaded-regression copse (RCRC) [10] for all still images and video frames of PaSC are now publicly available[5].

E. University of Technology, Sydney (UTS)

The approach taken by the group from UTS handles the rich variations in the video with a robust face representation which is modified from the approach presented by Ding et al. [8]. The approach features three-dimensional (3D) face pose normalization and two effective face descriptors; i.e., Dual-Cross Patterns (DCP) [7] and LPQ [20].

In detail, based on the eye coordinates and associated bounding face boxes provided by the competition, five facial feature points are further detected: two eye centers, the nose tip, and the two mouth corners. The five facial feature points in the two-dimensional (2D) image are first aligned with those of a generic 3D face model [21]. The textures of the 2D face are then mapped to the 3D model. Finally, a frontal face is rendered with the textured 3D model. The size of the rendered face is 156×130 pixels. Note, the horizontally flipped faces are also utilized.

Multi-scale DCP and LPQ descriptors were employed for feature extraction. The features were extracted from the left half face and the right half face, respectively. If one half of the face was severely occluded due to the pose variation, then we did not extract features from that half face. The DCP features were extracted at three scales, with parameters set at [2, 4], [4, 8], and [6, 12]. The LPQ features were extracted at six scales, with parameters set at 3, 5, 7, 9, 11, and 13. The half face images were divided into 12 x 6 pixel non-overlapping regions. The multi-scale DCP features were extracted and concatenated from the 72 regions and the same strategy was applied to the multi-scale LPQ features. Corresponding feature vectors of all faces in a video were averaged to obtain the face representation of the video. Therefore, there were four feature vectors utilized to represent a video; i.e., the averaged multi-scale DCP features of the left half face, the averaged multi-scale LPQ features of the left half face, the averaged multi-scale DCP features of the right half face, and the averaged multi-scale LPQ features of the right half face.

The dimensionality of each feature vector of the video was first reduced to 600 by principal component analysis (PCA). Then face matching was conducted using the PLDA [23] model. The similarity scores of the four classifiers were fused by the sum rule. Both the PCA and PLDA models were trained on the LFW database [12].

V. Results

The ROC curves for the control and handheld experiments are presented in Figure 2. Included in Figure 2 is performance of the baseline local region PCA (LRPCA) algorithm, which is part of the PaSC distribution. The verification rates at FAR=0.01 for the five participants on the control and handheld videos are noted on the ROC plot. Several things are evident from these results. First, recognition is harder on the handheld videos, and the difference between the control and handheld videos is generally not dramatic. Second, there is a wide range of performance in this evaluation, and the top performer is delivering verification rates at FAR=0.01 in the 0.5 to 0.6 range, which underscores the difficultly of the problem. The results on the handheld experiment are directly comparable to the results from the IJCB 2014 competition [4], where the top verification rate at FAR=0.01 was 0.26. The results for the IJCB competition were due in April 2014 and the due date for this competition was in November 2014. Therefore, in a six month period we see a doubling in verification performance for the handheld video.

VI. Covariate Analysis

How performance on the handheld video was influenced by a series of factors is summarized in Figure 3. This analysis repeats the analysis presented in Beveridge et al. [4] and Lee et al. [15] for the submissions in this competition. The

[5] https://sites.google.com/site/zhenhuaswebpage/home/pasc_landmarks

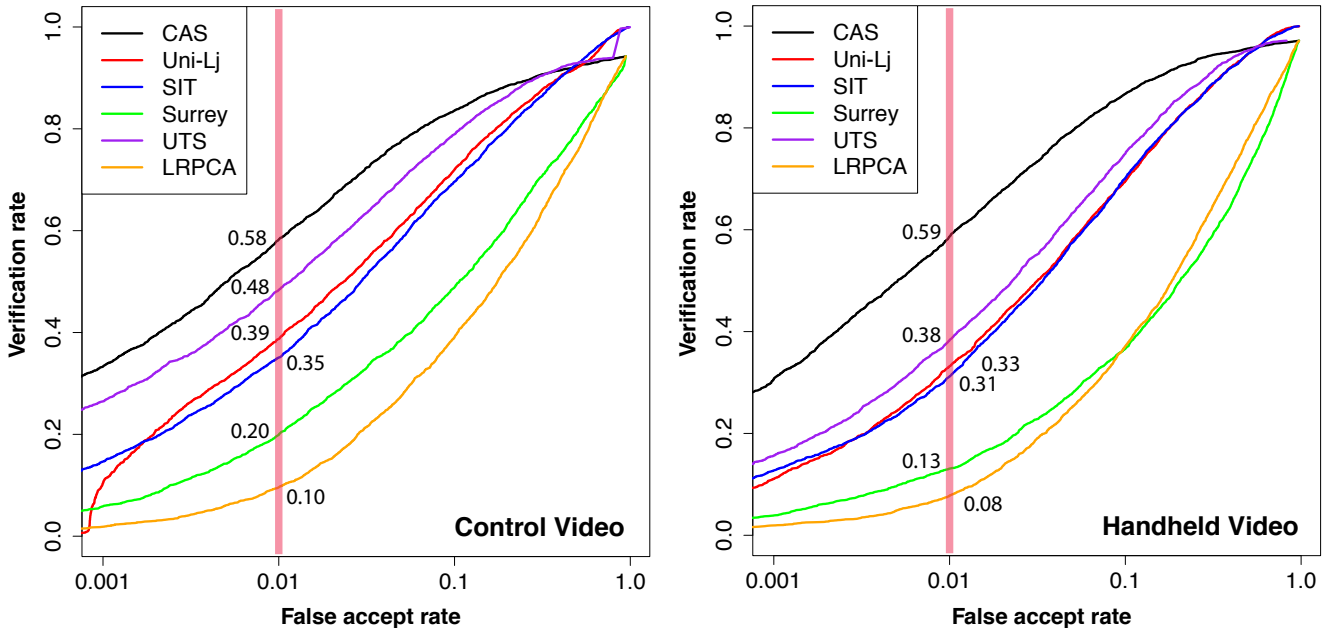

Fig. 2. ROC curves for the control and handheld video evaluations.

vertical axis in all cases is the verification rate at FAR=0.01. The first four plots from left to right represent changes in the face yaw, roll, size and detection confidence as measured by the PittPatt SDK 5.2.2 face detection algorithm. Cases are divided into three bins: small (S), medium (M) and large (L) for each factor. So, for example, average face size over each video is determined and then used to divide matching results into three equal sized bins, labeled S, M and L. Then the verification rate is reported for each bin and lines are drawn between to highlight trends. For face size, all three algorithms show a significant size effect, finding larger faces easier to recognize.

The next two plots investigate the role that environment plays in influencing recognition difficulty. There are three components to environment as expressed here: location, camera and action. Recall the summary in Table II. The first plot shows the environment/action pairs with the lowest verification rate on the left, an intermediate combination in the middle, and the combination with the highest verification rate on the right. The actual labels change by algorithm because not all algorithms found the same combinations easiest/hardest. The second plot is similar, but broken down by handheld sensor combinations. Because sensors were paired with locations and actions, it is not possible to fully separate the influence of camera and location/action, however it is clear from the Env(Act) and Sensor plots that both are playing a major role in influencing performance. So, for example, notice that going from the easiest environment/action to the hardest, all three algorithms see at least a doubling in verification rate. This is particularly noteworthy given the actual verification rates are in some cases shifted considerably in absolute terms.

The variation associated with different people, the SubID plot, shows a significant range of difficulty between the easiest versus the hardest sets of people. However, it is important to note when interpreting this finding that since the hardest versus easiest person distinction is made using the results of the algorithm itself, this only shows that there is wide variation for each algorithm and should not be confused with an actual analysis of which people are hard or easy and whether that distinction is stable between algorithms.

The last two plots indicate that videos with male subjects are consistently easier by a modest amount and also a modest improvement for Asian subjects relative to Caucasian. Both of these findings are consistent with previous face recognition covariate studies [3], [19]. While it is important to measure gender and race influence, it is also important to notice they are secondary factors in terms of importance relative to environment, sensor or subject variation.

VII. Conclusion

Person identification in video when the people are in motion and not attending to the camera is difficult. The Point-and-Shoot Challenge (PaSC) video data is an open dataset that allows the research community to test algorithms against a set of videos acquired under these challenging conditions. The control and handheld PaSC videos provided the basis for the Face and Gesture 2015 evaluation. When first released in 2013, the baseline algorithm provided by CSU achieved a verification rate of only 0.08 at FAR=0.1 on the PaSC handheld video. The best performance reported in 2013 was achieved using a commercial algorithm, the PittPatt SDK 5.2.2, and that algorithm achieved a verification rate of 0.38 at FAR=0.01. In the current evaluation, five labs from around the world participated and all exceeded the performance of the CSU baseline, most by a considerable amount. In addition, the University of Technology, Sydney, matched the performance of the PittPatt algorithm, and

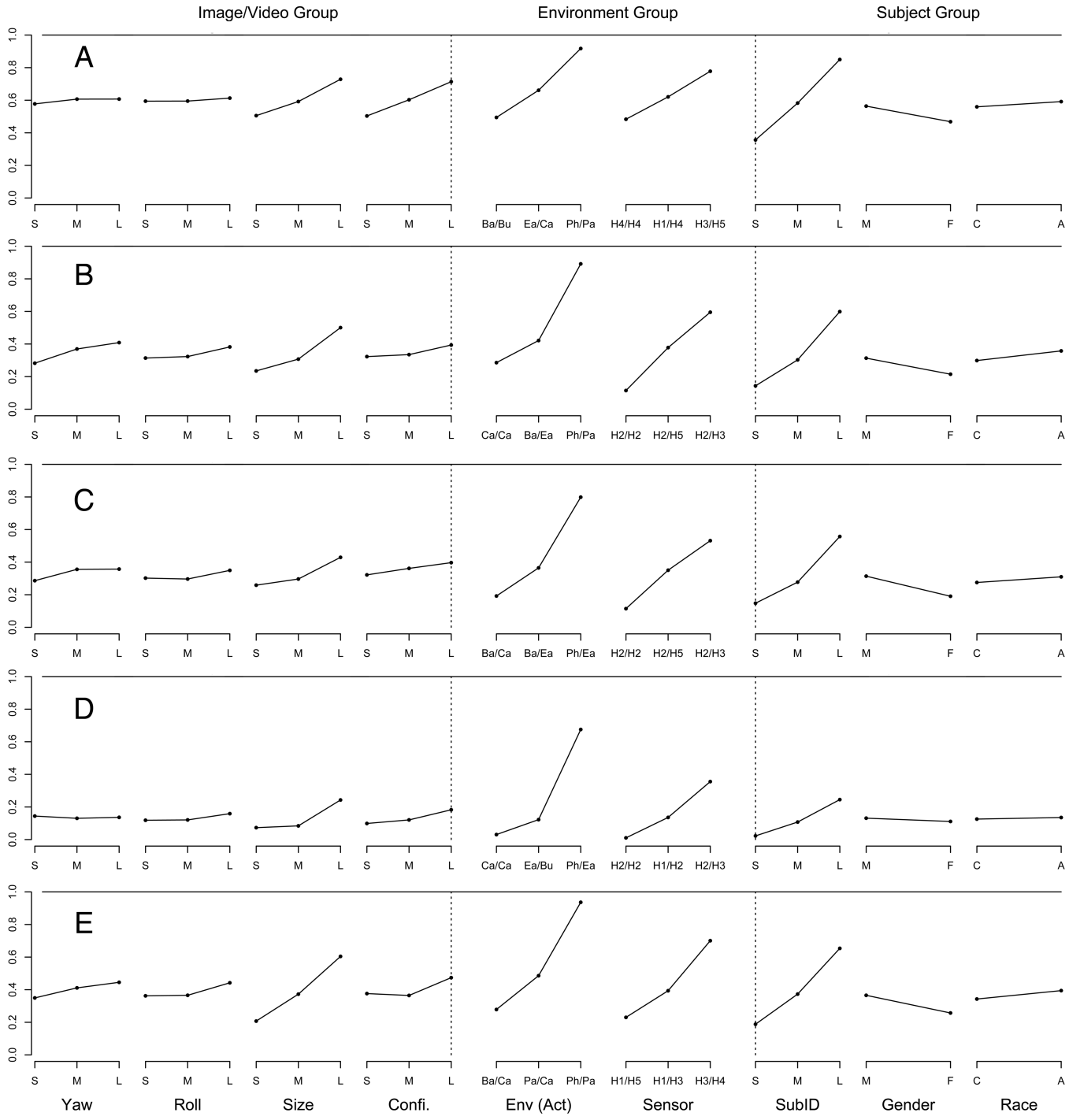

Fig. 3. Marginal analysis of verification rate change, shown on the vertical axis, conditioned on different factors using VR@FAR=0.01 across all five participants. The participants are identified by letter, A: Chinese Academy of Science (CAS), B: University of Ljubljana (Uni-Lj), C: Stevens Institute of Technology (SIT), D: University of Surrey (Surrey), E: University of Technology, Sydney (UTS).

the Chinese Academy of Science algorithm did far better, delivering a verification rate of 0.59 at FAR=0.01. This evaluation highlights the progress made in video person recognition, with performance increasing significantly since the introduction of the PaSC video, and with top performers now beating the initial high performance level established by the PittPatt algorithm.

Two additional findings are worth note. The first is the strong dependence on location and action. All algorithms showed a significant variation in performance level when evaluated on specific settings and actions. This result further corroborates a finding initially reported as part of the IJCB 2014 Handheld Video Face and Person Recognition Competition [4]. A new and surprising finding is that for the top performing algorithm there is little difference in verification performance between the control and handheld video: 0.58 for control and 0.59 for handheld. This runs counter to expectation, since the control video is higher resolution video, always taken with the same type of camera that was mounted on a tripod, hence a stable camera. The results for the other four participants shows the expected rise in performance between the handheld and control video. That the top performing algorithm does not show this dependence is both a credit to the algorithm and a suggestions of how much more remains to be done and understood about performance on video taken in unconstrained environments.

REFERENCES

[1] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Comput.*, 12(10):2385–2404, Oct. 2000.

[2] 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 *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pages 1–8, Sept 2013.

[3] J. R. Beveridge, G. H. Givens, P. J. Phillips, and B. A. Draper. Factors that influence algorithm performance in the face recognition grand challenge. *Computer Vision and Image Understanding*, 113(6):750 – 762, 2009.

[4] J. R. Beveridge, H. Zhang, P. Flynn, Y. Lee, V. E. Liong, J. Lu, M. Angeloni, T. Pereira, H. Li, G. Hua, V. Struc, J. K. V. Štruc, and a. J. P. J. Križaj. The IJCB 2014 PaSC Video Face and Person Recognition Competition. In *International Joint Conference on Biometrics*, September 2014.

[5] C. H. Chan, M. Tahir, J. Kittler, and M. Pietikäinen. Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(5):1164–1177, May 2013.

[6] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision – ECCV 2012*, volume 7574 of *Lecture Notes in Computer Science*, pages 566–579. Springer Berlin Heidelberg, 2012.

[7] C. Ding, J. Choi, D. Tao, and L. S. Davis. Multi-directional multi-level dual-cross patterns for robust face recognition. *arXiv preprint arXiv:1401.5311*, 2014.

[8] C. Ding, C. Xu, and D. Tao. Multi-task pose-invariant face recognition. *IEEE Transactions on Image Processing*, 2015.

[9] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[10] Z.-H. Feng, P. Huber, J. Kittler, W. Christmas, and X.-J. Wu. Random cascaded-regression copse for robust facial landmark detection. *Signal Processing Letters, IEEE*, 22(1):76–80, January 2015.

[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 *The 6th IAPR International Conference on Biometrics*, June 2013.

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[13] Z. Huang, R. Wang, S. Shan, and X. Chen. Hybrid Euclidean-and-Riemannian Metric Learning for Image Set Classification. In *Proceedings of the 12th Asian Conference on Computer Vision (ACCV 2014)*, Singapore, November 2014.

[14] Z. Huang, R. Wang, S. Shan, and X. Chen. Learning euclidean-to-riemannian metric for point-to-set classification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1677–1684, June 2014.

[15] Y. Lee, P. J. Phillips, J. J. Filliben, J. R. Beveridge, and H. Zhang. Generalizing face quality and factor measures to video. In *Proceedings of the 2014 International Joint Conference on Biometrics*, September 2014.

[16] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3499–3506. IEEE, 2013.

[17] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt. Eigen-Pep for Video Face Recognition. In *Proceedings of the 12th Asian Conference on Computer Vision (ACCV 2014)*, 2104.

[18] P. Li, Y. Fu, U. Mohammed, J. Elder, and S. J. Prince. Probablistic models for inference about identity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):144–157, 2012.

[19] Y. M. Lui, D. Bolme, B. Draper, J. Beveridge, G. Givens, and P. Phillips. A meta-analysis of face recognition covariates. In *Biometrics: Theory, Applications, and Systems, 2009. BTAS '09. IEEE 3rd International Conference on*, pages 1–8, Sept 2009.

[20] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In *Image and signal processing*, pages 236–243. Springer, 2008.

[21] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance*, pages 296–301, 2009.

[22] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 947–954 vol. 1, June 2005.

[23] S. J. Prince and J. H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1–8, 2007.

[24] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR 2014 Proceedings*, 2014.

[25] J. K. V. Štruc and S. Dobrišek. MODEST face recognition. In *International Workshop on Biometrics and Forensics (IWBF'15) (under review)*, 2015.

[26] V. Štruc and N. Pavešić. The Complete Gabor-Fisher Classifier for Robust Face Recognition. *EURASIP Journal on Advances in Signal Processing*, 2010(1), 2010.

[27] N.-S. Vu and A. Caplier. Illumination-robust face recognition using retina modeling. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 3289–3292, November 2009.

[28] R. Wang, H. Guo, L. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

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

[30] L. Wolf, T. Hassner, and Y. Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(10):1978–1990, Oct 2011.

[31] X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum. Finding celebrities in billions of web images. *Multimedia, IEEE Transactions on*, 14(4):995–1007, 2012.