NISTIR 8004

Identifying Face Quality and Factor Measures for Video

Yooyoung Lee
P. Jonathon Phillips
James J. Filliben
J. Ross Beveridge
Hao Zhang

http://dx.doi.org/10.6028/NIST.IR.8004

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 8004

Identifying Face Quality and Factor Measures for Video

Yooyoung Lee
P. Jonathon Phillips
Information Access Division
Information Technology Laboratory

James J. Filliben
Statistical Engineering Division
Information Technology Laboratory

J. Ross Beveridge
Hao Zhang
Computer Science Department
Colorado State University
Fort Collins, CO

http://dx.doi.org/10.6028/NIST.IR.8004

May 2014

U.S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Patrick D. Gallagher, Under Secretary for Standards and Technology and Director

Identifying Face Quality and Factor Measures for Video

Yooyoung Lee, P. Jonathon Phillips, James J. Filliben, J. Ross Beveridge, Hao Zhang

National Institute of Standards and Technology, Information Technology Laboratory
100 Bureau Drive, Gaithersburg, MD 20899, USA
Colorado State University, the Computer Science Department, Fort Collins, CO 80523, USA
{yooyoung, jonathon, filliben}@nist.gov, {ross, zhangh}@cs.colostate.edu

Abstract

This paper identifies important factors for face recognition algorithm performance in video. The goal of this study is to understand key factors that affect algorithm performance and to characterize the algorithm performance. We evaluate four factor metrics for a single video as well as two comparative metrics for pairs of videos. This study carried out an investigation of the effect of nine factors on three algorithms using the Point-and-Shoot Challenge (PaSC) video dataset. These factors can be categorized into three groups: 1) image/video (pose yaw, pose roll, face size, and face detection confidence); 2) environment (environmental condition with person's activity and sensor model); and 3) subject (subject ID, gender, and race). For video-based face recognition, the analysis shows that the distribution-based methods were generally more effective in quantifying factor values. For predicting face recognition performance in a video, we observed that face detection confidence and face size serve as potentially useful quality measure metrics. We also find that male faces are easier to identify than female faces, and Asians are easier than Caucasians. Further, on the PaSC video dataset, the performance of face recognition algorithms are primarily driven by environment and sensor factors.

Keywords: face recognition, sensitivity analysis, factor analysis, biometrics, forensics, video surveillance

1. Introduction

As a discipline, face recognition from video has grown due to the broad range of videos now being taken "anytime and anywhere" (e.g., webcam, CCTV, and mobile devices including cell phone.) Recognizing a person's face from a wide spectrum of video domains presents significant challenges. Unlike static face images, a sequence of video contains multiple frames of the same face and it can acquire a face from multiple angles, illuminations, and expressions [1]. There are also numerous factors that can affect the performance of video-based face recognition algorithms. Beyond facial characteristics, such factors include camera types, location/background, illumination, subject action (e.g., pose and expression), distance, body-visible, algorithmic parameters, etc.

Factor effect analysis (sensitivity analysis) is the study of how the output of an algorithm is affected by different inputs. A number of studies on factors that affect algorithm performance have been conducted [2][3][4][5]. In face recognition from still images, it is well understood how to assess the impact of factors on algorithm performance. However, generalizing factors from still images to video sequences has not been studied. For example, stating the pose or size of a face in a still image is straightforward, but what is the pose or size of a face in a video sequence?

One major contribution of our study is that while these prior studies focused on the factor effects of face

recognition in still-based face images, our study generalizes factor value measures to video sequences. The generalization is based on the distribution of factors estimated from each frame. We conduct a sensitivity analysis based on our new methodology. We demonstrate the analysis methodology with nine factors on three algorithms using the Point-and-Shoot Challenge (PaSC) video dataset [6]. The three algorithms are: the open-source Local Region PCA (LRPCA) [7], the commercial algorithm PittPatt [8], and the Principal Angle (PA) algorithm [9][10] developed by Colorado State University (CSU), respectively. This study provides a comparative characterization of the algorithms themselves, and delivers a ranking (and understanding) of key factors that affect algorithm performance.

Specifically, this work addresses the following questions: (i) What is the best manner to extend factor measures from still to video imagery? (ii) How do we quantify factors for video? (iii) Are such video factors suitable for sensitivity analysis? What are the relative important factors that affect the performance of face recognition algorithm in video? (iv) Are conclusions about factors robustly true over multiple algorithms? (v) Can our analysis methods be used to derive quality measure metrics for predicting general performance of face recognition algorithms?

The impacts/contributions of this paper are as follows:

- A determination of the most important factors (and interactions) for face recognition in video;
- A baseline methodology framework for quantifying/generalizing factor measures for video;
- A characterization of current face recognition technology that provides guidance on how in the real-world we can improve existing algorithms and develop new algorithms;
- A leveraging of the existing factor studies on stills to the more challenging videos; and
- A useful methodology for forming quality measure metrics for predicting algorithm performance—which improves on existing image quality measures by rigorously investigating and identifying key factors that affect algorithm performance.

The examination of the effects of factors on the performance of biometric algorithms in the video domain is a relatively new research approach, and it extends the important preexisting factor studies on stills [2][3][5]. We believe that the methodology demonstrated herein can be applied to other video-based biometric systems (e.g., gait, body, iris). The methods could also be utilized in related law enforcement applications, for example, in providing a means to assess the relative importance of facial features (using a systematic statistical analysis), which would be an invaluable tool for security and forensics.

2. Overview of Methodology and Results

This section provides a brief overview of our methodology and results. The nine factors used in our experiment can be categorized into three groups: 1) image/video (pose yaw, pose roll, face size, and face detection confidence); 2) environment (environmental condition with person's activity and sensor/camera model); and 3) subject (subject identify, gender, and race). We first examine the factor measures within the image group. Figure 1 shows a summary of the approach and key findings for generalizing/quantifying factor values for video.

We also investigate the cross-environment and cross-sensor analysis for factors of the environment group. For subject group factors, we use graphical analyses to characterize the factor effects on algorithm performance. Finally, for the nine factors, we carry out a sensitivity analysis and provide a ranked list of factors based on their relative importance to the performance of face recognition algorithms in video. A summary of the sensitivity analysis is illustrated in Figure 2.

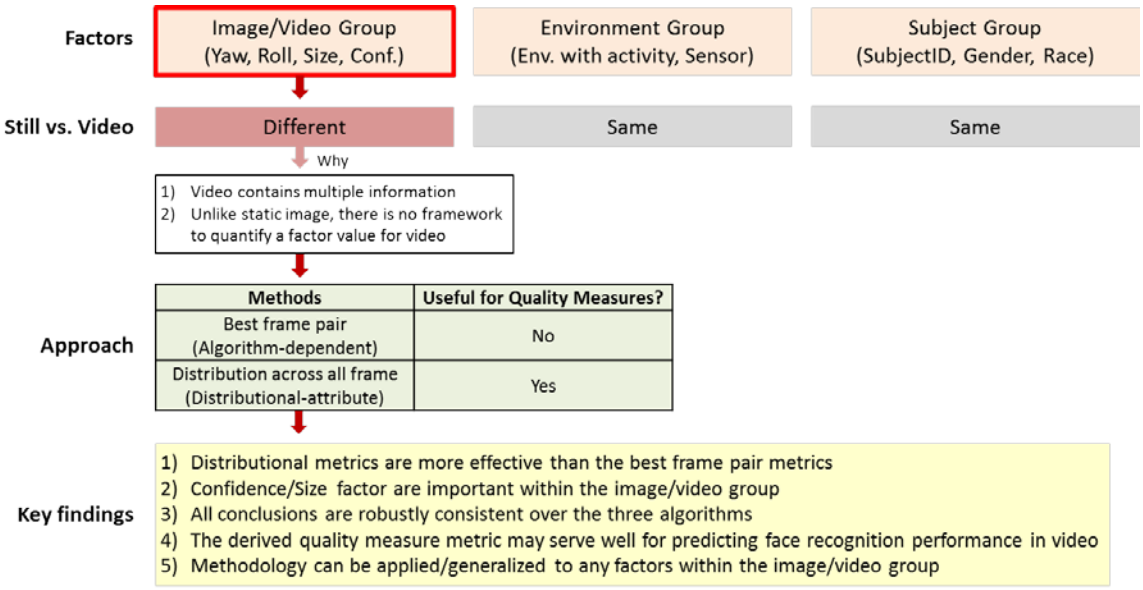

Figure 1. Summary of factor generalization for video

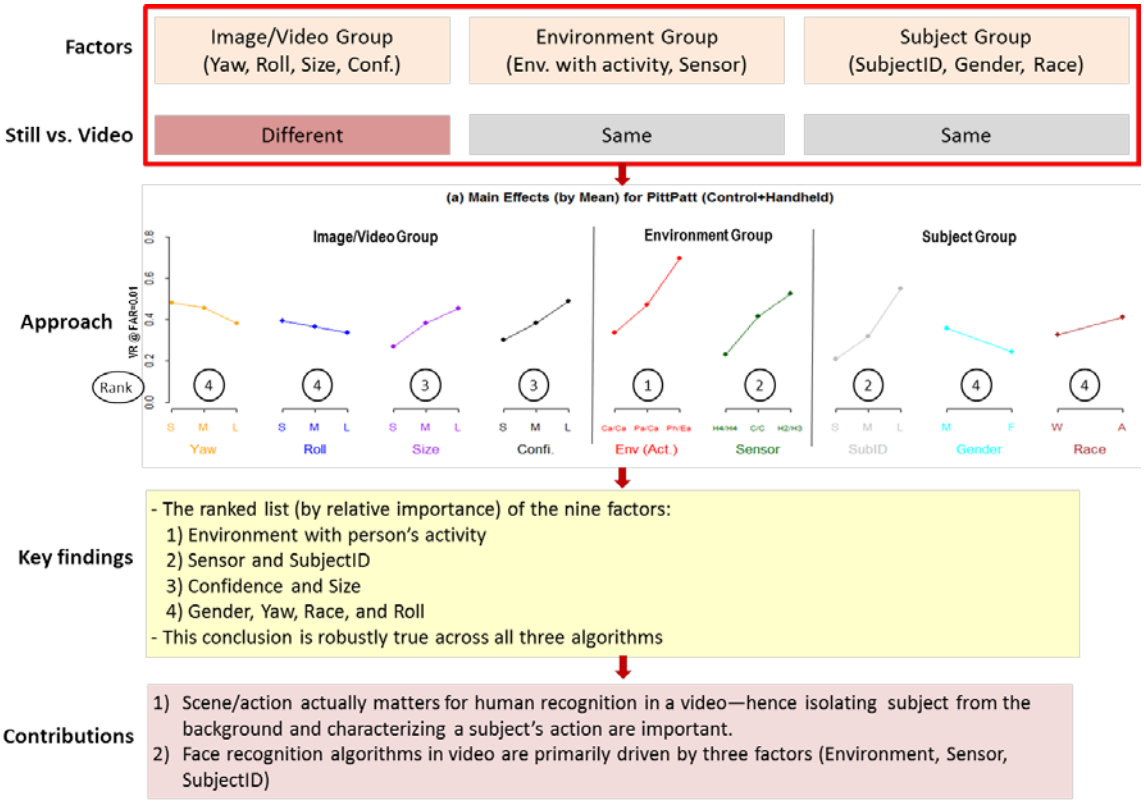

Figure 2. Summary of the sensitivity analysis for the nine factors; the x-axis is the factor name and its level settings, and the y-axis is the mean response (VR at FAR=0.01); the number marked in circle indicates the ranking of the factor effects

In the Main Effects plot of Figure 2 (PittPatt algorithm case), the x-axis lists the factor name and its discretized level settings, and the y-axis is the mean response (verification rate at a false accept rate = 0.01) for a given factor and level. The number marked in circle indicates the ranking of the factor effects on algorithm performance. Over the nine factors, Environment with person's activity had the most effect on performance in video, followed by Sensor and SubjectID—these conclusions were true for each of the three algorithms. We thus observe that scene/action actually matters for face (human) recognition—this implies the importance of isolating a subject from the background and of characterizing a subject's action.

3. Dataset

Face images taken in controlled environments (e.g., mugshot) have a high verification rate [11]. These days, however, face images are often taken via mobile devices (including cell phones), which presents a number of challenges in terms of individual identification.

To address these issues, the Point-and-Shoot Challenge (PaSC) dataset was recently introduced by Beveridge et al. [6]. Although the dataset includes both still and video imagery, our study focuses on the video sequences to demonstrate the analysis methodology and results.

PaSC contains videos that were captured by six different sensors, in six different environmental conditions, with subject movement at a distance. Figure 3 shows an example of video frames with their corresponding environment with person's activity. The sensor type is either a handheld or a tripod-mounted camera (namely, a control camera). Unfortunately, the five sensors ("handheld") are indistinguishably confounded with the six environments, which means that sensor and environment effects are inextricably mixed. On the other hand, one of the sensors ("control") has collected videos over all six environments which allows us to have a narrower conclusion of environment effect.

The number of videos for handheld (1401) and control (1401) are evenly distributed. The environment factor in this study includes the subject's different activities at different locations, indoor/outdoor, different illuminations, etc. The details regarding environment and sensor factors and their settings are discussed in Section 4.2.

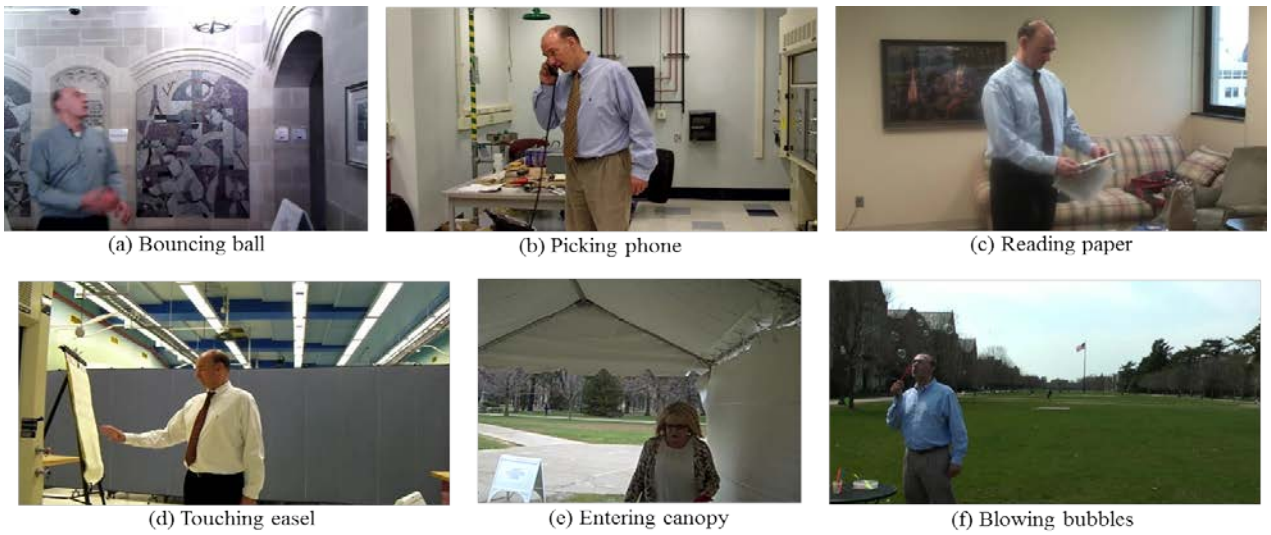

Figure 3. Video frame examples from the PaSC video dataset

(a) FVRT2006

(b) Sub-images from the PaSC video dataset

Figure 4. Comparison of (a) traditional face image and (b) challenging face images from PaSC

Figure 4 shows example images from the FVRT2006 dataset [12] and the PaSC dataset—note the image quality differences between (a) traditional still-based face images, and (b) near real-world environment video-based face images. These differences show the challenge of recognizing a face in less-than-optimal environments, while coping with low video quality. Our work thus extrapolates still-based face recognition to more challenging video-based face recognition.

4. Method

Measuring characteristics of an algorithm assist in the understanding and improvement of already-developed algorithms as well as future algorithms [13]. In this section, we introduce examples of factor measures for characterizing face recognition algorithms in a video. The main goal of the analysis is to provide researchers insight into which factors are more important and less important to the performance of video-based face recognition. In addition, a systematic statistical analysis provides insight on where scientists could focus their near-term and long-term research efforts [14], i.e. what to consider when developing future algorithms to overcome even further challenges.

4.1 Algorithm and Performance

To demonstrate the analysis methodology and its robustness, we used three face recognition algorithms: Local Region PCA (LRPCA) [7], Pittsburgh Pattern Recognition (PittPatt) [6], and Principal Angle (PA) [9][10]. LRPCA is an open source algorithm developed as a baseline for face recognition; PittPatt is a black-box commercial system; and PA is an image-set to image-set matching algorithm.

Algorithm performance is normally assessed by the comparison of video pairs, which we call the query (input) and the target (reference) video. The first step in video-based face recognition is to detect/extract a face within each frame. The PittPatt detection algorithm was utilized throughout to extract face sub-images within each frame in a video. A face was not detected in every frame; therefore, not every frame in a video contributed a face sub-image. For some videos, no face was detected; we thus omitted 42 out of 1401 control videos and 20 out of 1401 handheld videos in this study. Our analysis is based on the 2740 videos where a face was detected in at least one frame. On average, a face was detected in ≈100 frames in each video.

For LRPCA- and PittPatt-based face recognition algorithm for video, the first step is to compare all face-present frames within a query video to all frames in a target video. After ranking all frame pairs based on the scores, the

frame pair with the highest similarity score [6] is then selected. In PA, all face sub-images from a video are preprocessed (1) by normalizing them to 32x42 pixels, and (2) by aligning the different face angles via flipping the non-frontal face images to face the same direction. To compute a similarity score for PA, a random subset of frames from each query and target video is selected, and then the principle angle is calculated between the two images sets. The image set consists of the normalized face regions randomly selected.

The major difference between the (LRPCA, PittPatt) algorithms and the PA algorithm is that LRPCA and PittPatt ultimately use information from one frame pair to assess similarity between two videos, while the PA algorithm uses information of sets of images extracted from randomly sampled frames from a video.

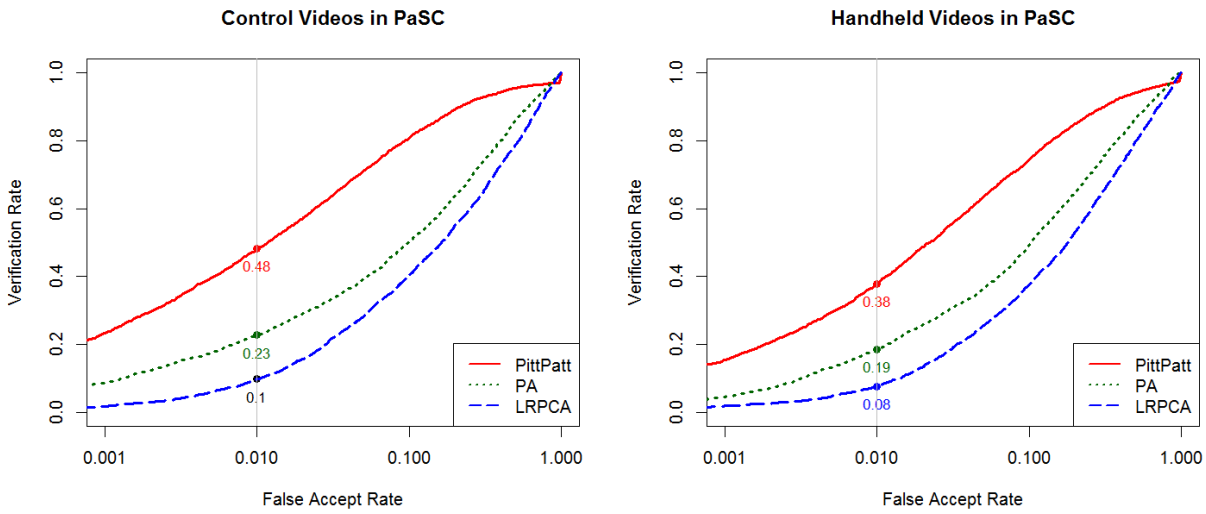

Figure 5. Performance of three algorithms on PaSC for control and handheld videos

The total number of similarity scores calculated for all possible pairs is approximately 1.87M (= 0.92M+0.95M) for both control (1359) and handheld (1381) videos. Figure 5 shows the performance of the three algorithms with the PaSC data for control (left) and handheld (right) videos. For the control data, the PittPatt algorithm has 48 % verification rate (VR) at a false accept rate (FAR) = 0.01, 23 % for PA, and 10 % for LRPCA. For the handheld sensor data, the VR at a FAR =0.01 is 38 % for PittPatt, 19 % for PA, and 8 % for LRPCA. For both control and handheld videos, PittPatt has the highest performance followed by PA and LRPCA.

4.2 Factors

While there are many factors that can affect algorithm performance on face recognition in a video, our study focuses on nine: 1) face pose yaw, 2) face pose roll, 3) face size, 4) face detection confidence, 5) environment, 6) sensor (camera) model, 7) subject ID (individual), 8) gender, and 9) race. These factors can be categorized into local and global. A local factor is a measurable (typically continuous and ordinal) quantity which describes some particular feature, attribute, or characteristics of the image/video, and which can possibly be predictive of the performance of recognition algorithms. A global factor is a non-measurable (typically discrete and categorical) variable which describes some aspect of the image/video. For video, local factor values can change frame-by-frame, whereas global variables normally remain constant over the set of frames.

In our case, the four factors yaw, roll, size, and confidence (as estimated by PittPatt SDK) are local, while the remaining five factors (environment, sensor, subjectID, gender, and race) are global.

We choose these nine factors because they have greater immediate research interest (among the authors), and

because they serve as a useful subset to demonstrate our analysis methodology. The nine factors are described in detail as follows:

1) *Face pose yaw (Yaw)*
   The pose yaw is the horizontal face rotational angle in degrees to the left and right. The range of the angles is approximately [-40^0 to 40^0].
2) *Face pose roll (Roll)*
   The pose roll is the angle of rolling/tilting the head to the shoulder. The PaSC video dataset has a range of [-30^0 to 30^0]—outliers exist for this factor (less than 0.001 %), which we omitted from our study.
3) *Face size (Size)*
   The face size is an approximation of face size (resolution) as reported by the PittPatt SDK [8]. The raw units for this factor range are 0 to 16 on a logarithmic scale.
4) *Face detection confidence (Confidence)*
   The face detection confidence is the PittPatt SDK's self-assessment of correct face detection [8] with a range of 0 to 10—where larger values indicate a higher confidence for correct face detection.
5) *Environment with Activity (Environment)*
   The PaSC video dataset has six different environment settings as shown in Table 1.

Table 1. Environment information

Code	Location
Ball	Indoor
Phone	Indoor
Paper	Indoor
Easel	Indoor
Canopy	Outdoor
Bubble	Outdoor

   This factor is associated with multiple factors. Four key contributing factors are subject activity, background, indoor/outdoor, and lighting condition. Since it is difficult to separate these variables as individual factors, we combine these variables into a single environment factor. For example, in videos taken of the scene Ball, a person walks and bounces a ball indoors in front of a complex background. In the Phone scene, a person walks towards a phone and picks it up. For the Paper scene, a person walks while reading a paper. For the Easel scene, a person walks toward an easel, touches it, and leaves the scene. For the Canopy scene, a person enters a canopy from the outside towards the camera. Lastly, for the Bubble scene, a person walks towards a table, blows bubbles outside, and finally leaves the scene.
6) *Sensor*
   The PaSC video dataset was collected by six different cameras as shown in Table 2.

Table 2. Sensor (camera) information

Code	Type	Sensor Model	Size	Env. Code
C	Control	Panasonic HD700	1920x1080	All
H1	Handheld	Flip Mino F360B	640x480	Canopy
H2	Handheld	Kodak Zi8	1280x720	Phone, Canopy
H3	Handheld	Samsung M. CAM	1280x720	Paper
H4	Handheld	Sanyo Xacti	1280x720	Easel, Bubble
H5	Handheld	Nexus Phone	720x480	Ball

   The Panasonic HD HS 700 (coded as C) is the only stationary control camera, while the remaining entries are a handheld camera. The control camera was used in all six environments. The PaSC data was collected with five handheld cameras labeled H1 through H5. The labels were chosen to focus attention on the effects of different sensor and not on the make of the cameras. For the handheld cameras (H2 and H4), data was collected in two different environments, and for the remaining handheld cameras (H1, H3, H5), data was collected in only one

environment. Note that the PaSC dataset has the property that the Environment and Sensor factors are confounded and unbalanced. For example, one cannot separate the environment Ball effect from the sensor H5 effect, because the sensor H5 was the only sensor used for the environment Ball. The use of the control sensor C for each environment helps in this regard, but a more balanced design/dataset would have considerably enhanced the estimation/separation of the effects and the robustness of the environment and sensor factor conclusions.

7) *SubjectID (Individual)*
   The PaSC data are from 265 individuals. The average number of videos per person is approximately 22.
8) *Gender*
   Out of 265 subjects, the gender is fairly balanced between males (55 %) and females (45 %).
9) *Race*
   The PaSC videos contain five different races: Caucasian, Asian-Pacific, Black-American, Hispanic, and Unknown. The population is mainly driven by Caucasians (81 %) followed by Asian-Pacific (12 %). Black-Americans account for 2 % of the subjects, and Hispanics account for 1 %. The remaining 5 % are of unknown race. Some races don't have a sufficient number of subjects for evaluation; we thus include only Caucasian (White) and Asian-pacific (Asian) in our race study.

The nine factors can be divided by three groups; 1) image/video, 2) environment, and 3) subject. The four factors (Yaw, Roll, Size, and Confidence) are the image/video group which contains local variables. The environment group consists of two factors (Environment and Sensor) and the subject group consists of the three factors (SubjectID, Gender, and Race). These five factors are global variables.

4.3 Factor Measures

For the given nine factors, in this study, we have constructed metrics based on only those factors which are continuous, ordinal, and can change on a frame-by-frame basis, namely, Yaw, Roll, Size, and Confidence. These factors can potentially serve as predictors for biometric algorithm performance. We first describe how these factor metrics apply to still images and then discuss their extension to video.

A. Still image

Let q (query) and t (target) be two still face images. Given q and t, an algorithm A returns a similarity score $S_A(q,t)$, where a high score indicates that the subjects from the two face images are likely the same person.

Let X_k denote factor k from a set of factors under consideration. For our study, X_k is Yaw, Roll, Size, and Confidence. $X_k(q)$ and $X_k(t)$ are factor k values from images q and t, respectively. A goal of this study is to characterize and predict algorithm performance $S_A(q,t)$ by $X_k(q)$ and $X_k(t)$.

Since a similarity score $S_A(q,t)$ is based on a pair of images, the similarity score not only is related to $X_k(q)$ and $X_k(t)$ individually, but also to $X_k(q)$ and $X_k(t)$ jointly. To measure such behavior of a pair of $X_k(q)$ and $X_k(t)$, we construct a comparative metric. The comparative metric is a function of image factors that produces a quantitative evaluation of the similarity of the two images with respect to that factor. We denote the comparative metric as:

$$M_k(q,t) = g(X_k(q), X_k(t))$$

To return a comparative score $M_k(q,t)$, we use examples of g as follows:

1) absolute difference (Δ): $|X_k(q) - X_k(t)|$
2) extremum (max/min): $|max(X_k(q), X_k(t))|$ or $min(X_k(q), X_k(t))$

For absolute difference (Δ), a near-zero comparative score of Δ indicates that factor k values are near-identical between images q and t. This absolute difference is used for all four factors. For example, for Yaw, a face pose direction from images q and t is near equivalent if Δ is near zero.

Before discussing the extremum approach, note that two images are considered "ideal" for comparison if 1) both Yaw values are near-frontal (near-zero angle), 2) both Roll values are near-frontal, 3) both Size values are large (higher resolution is better than smaller resolution), and 4) both Confidence values are large (higher face detection

confidence is better than smaller confidence). In this light, the extremum (max/min) approach provides the poorer value out of the two images. The choice of max and min depends on the factor; max is used for factors Yaw and Roll, and min for factors Size and Confidence. A small max and a large min indicate that factor k between images q and t has the ideal factor characteristics.

B. Extension to video

This section describes how we extend still-based algorithms to video and encompass X_k from still to video.

The straightforward method for video is that the input to an algorithm are two videos q and t, and the output is a similarity score $S_A(q,t)$. Let $q = \{q_1, q_2, \dots, q_Q\}$ frames and $t = \{t_1, t_2, \dots, t_T\}$ frames be a set of face extracted images from the query and target video, respectively. To extend still-based algorithms to video, all frame-to-frame pairs (q_i, t_j) provide a similarity score:

$$S_A(q,t) = f(S_A(q_i, t_i))$$

Usual f approaches are max, medium, mean, and 90th percentile. Let frame q^* and t^* be the best frame pair such that:

$$S_A(q,t) = S_A(q^*, t^*)$$

For a specific A algorithm, the similarity score $S_A(q,t)$ is thus provided by that particular frame pair (q^*, t^*), where the similarity score was obtained by the f function. For example, LRPCA and PittPatt are still-based face recognition algorithms. To return a similarity score for videos q and t, LRPCA uses the max rule and PittPatt uses the 90th percentile rule [6].

To demonstrate extending X_k from still to video for face recognition, we look at four different metrics (namely, single-video factor metric); 1) LRPCA best frame pair, 2) PittPatt best frame pair, 3) mean, and 4) sum of mean and standard deviation. The LRPCA and Pittpatt best frame pair are algorithm-dependent metrics, while the other two (mean and sum of mean and standard deviation) are distribution-attribute metrics.

1) Algorithm-dependent

   For a single video, the algorithm-dependent metrics provide a factor value from one frame pair out of all frame-to-frame pairs (query video and target video) used by a specific algorithm.

   To extend X_k from still to video, the factor values $X_k(q)$ and $X_k(t)$ are computed from corresponding frame q^* for the query video and t^* for the target video. The best frame pair method is:

   $$X_k(q) = X_{k,A}(q^*) \ \textit{and} \ X_k(t) = X_{k,A}(t^*)$$

   where A is an algorithm for face recognition, q^* and t^* are particular frame pair provided by A algorithm. For example, the best frame pair for the LRPCA algorithm, we denote the factor Yaw values as $Yaw_{lrpca}(q^*)$ and $Yaw_{lrpca}(t^*)$ and for the PittPatt algorithm as $Yaw_{pitt}(q^*)$ and $Yaw_{pitt}(t^*)$.

2) Distribution-attribute

   The distribution-attribute provides a factor value that is derived from the distribution of all frames within a video and we thus define as:

   $$X_k(q) = h(q_1, q_2, \dots, q_Q) \ \textit{and} \ X_k(t) = h(t_1, t_2, \dots, t_T)$$

   Examples of h are mean (μ) and sum of mean and standard deviation ($\mu + \sigma$). μ is location-based metric and $\mu + \sigma$ combines the location-based metric and the spread-sensitive metric.

   The comparative metric for videos q and t is defined as:

   $$M_k(q,t) = g(h(q_1, q_2, \dots, q_Q), h(t_1, t_2, \dots, t_T))$$

   where the g function for videos is the same as the still cases.

For each of these four metrics, we shall consider two functions g (Δ and max/min)—this thus leads to a total of eight metrics for each factor (See Table 3 for further details.)

5. Analyses and Results

This section discusses the analysis methods and results for the nine factors for each of the three algorithms (PittPatt, PA, and LRPCA). We first examine the eight factor metrics (four single-video times two comparative metrics) for factors Yaw, Roll, Size, and Confidence within the image/video group. Second, we investigate the cross-domain analysis for the environment group factors (Environment and Sensor). Third, we present results for the subject group factors (SubjectID, Gender, and Race). Finally, for the nine factors, we carry out a sensitivity analysis and provide a ranked list of factors based on their relative importance to the performance of face recognition algorithms in video.

5.1 Yaw, Roll, Size, and Confidence Analysis

The factor metrics introduced in Section 4.3 can be applied to only the local factors since these factor values are measurable and change on a frame-by-frame basis. We choose the four single-video metrics ($lrpca_{best}$, $pittpatt_{best}$, μ, and $\mu + \sigma$) and the two comparative metrics (Δ and min/max) for our experiment. For each of the four local factors, Table 3 summarizes the 8 (4 x 2) metrics.

Table 3. Summary of factor metrics for each of the four factors (Yaw, Roll, Size, and Confidence)

Factors	$M_k(q,t)$	Algorithm-dependent		Distribution-attribute	
		LRPCA Best Frame Pair ($lrpca_{best}$)	PittPatt Best Frame Pair ($pittpatt_{best}$)	Mean (μ)	Mean+SD ($\mu + \sigma$)
Yaw	ΔY	ΔY_{lrpca}	ΔY_{pitt}	ΔY_{μ}	$\Delta Y_{\mu+\sigma}$
	$maxY$	$maxY_{lrpca}$	$maxY_{pitt}$	$maxY_{\mu}$	$maxY_{\mu+\sigma}$
Roll	ΔR	ΔR_{lrpca}	ΔR_{pitt}	ΔR_{μ}	$\Delta R_{\mu+\sigma}$
	$maxR$	$maxR_{lrpca}$	$maxR_{pitt}$	$maxR_{\mu}$	$maxR_{\mu+\sigma}$
Size	ΔS	ΔS_{lrpca}	ΔS_{pitt}	ΔS_{μ}	$\Delta S_{\mu+\sigma}$
	$minS$	$minS_{lrpca}$	$minS_{pitt}$	$minS_{\mu}$	$minS_{\mu+\sigma}$
Confidence	ΔC	ΔC_{lrpca}	ΔC_{pitt}	ΔC_{μ}	$\Delta C_{\mu+\sigma}$
	$minC$	$minC_{lrpca}$	$minC_{pitt}$	$minC_{\mu}$	$minC_{\mu+\sigma}$

In this section, we analyze how these four factors affect algorithm performance. We are primarily interested in answering the following questions;

1) How do the suggested eight metrics relate to algorithm performance?
2) How do the "algorithm-dependent" metrics and the "distribution-attribute" metrics affect algorithm matching scores?
3) Are these conclusions robustly true for all three algorithms?
4) Do these metrics serve as quality measure metrics for predicting the performance of face recognition for the PaSC videos?

We first examine the histogram-scatter plots shown in Figure 6. The first set of plots depicts the relationship between the PittPatt matching scores and the factor values calculated by PittPatt best frame pair metric. The second set depicts the relationship between the PA matching score and mean metric. The third set depicts the relationship between the LRPCA matching scores and the LRPCA best frame pair metric. The horizontal axis shows the eight cases with comparative metrics "difference (Δ)" and "extremum (max/min)" for each of the four factors—see the different shape of histogram distributions between the "difference" and "extremum" metric. The vertical axis for the scatter plot is similarity (matching) scores in a total of 5986 genuine video pairs. The third row shows correlation coefficients (%) between factor metrics and matching scores. It is known from the iris-based biometrics study [12] that classification accuracy is primarily driven by matching scores, not non-matching scores. We observed the same for video-based face recognition in our analysis (not shown)—that is, the shape and location of the non-matching

score distribution are more steady (contribute less to classification) across different image quality conditions. Note that the histogram distribution shapes for each separate control and handheld dataset are relatively similar to the shapes for the combined (control+handheld) video data; hence we chose to show the results for all (control+handheld) videos only.

Figure 6 (a) shows that, for the eight metrics for the PittPatt algorithm, $minC_{pitt}$ has the highest correlation (43 %) for all (control+handheld) videos--this observation is consistent for control-only (59 %) and handheld-only (41 %).

Figure 6 also shows that the Confidence factor with the comparative metric min (namely, $minC$) has the highest correlation with matching scores for all three algorithms (PittPatt: 43 %, PA: 49 %, LRPCA: 53 %). This indicates the important result that the selection of the poorer value for the Confidence factor between the query and target video has higher correlation with algorithm performance than do the other comparative metrics.

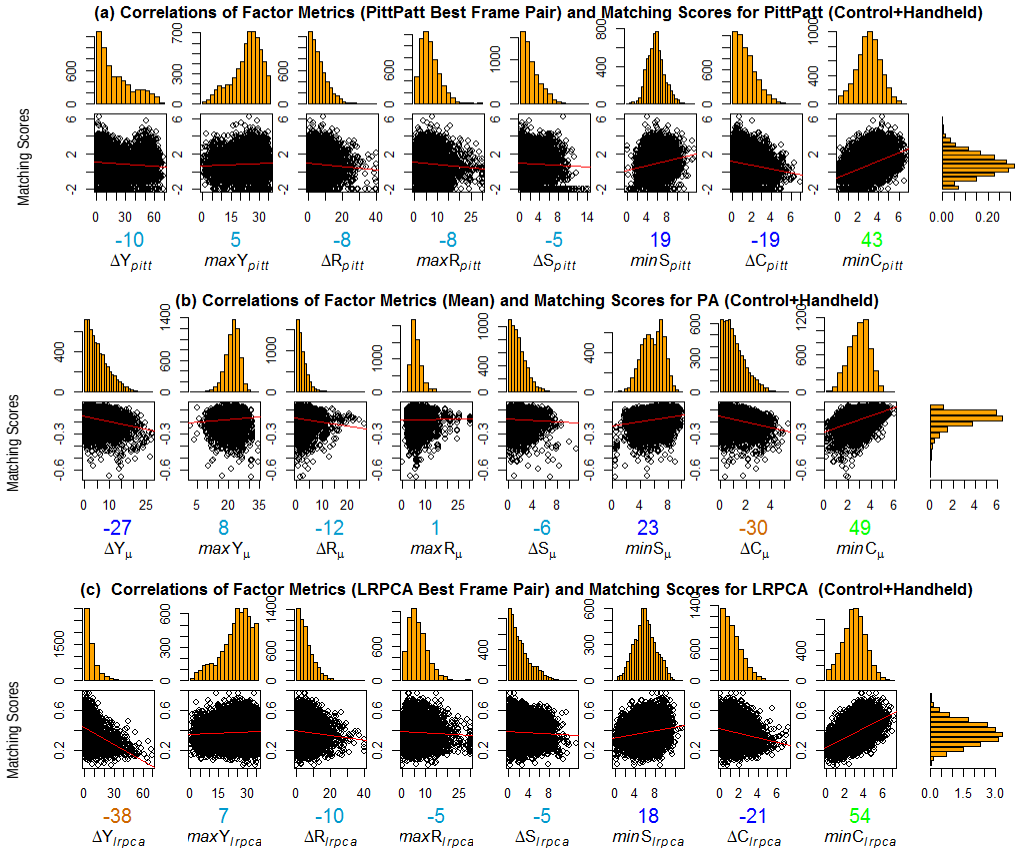

Figure 6. Correlations of the factor metrics and matching scores for factors Yaw, Roll, Size, and Confidence on the PaSC (control+handheld) video dataset; (a) PittPatt, (b) PA, and (c) LRPCA

Table 4 is a summary of the correlations for each of the three algorithms for all metrics and all data (control+handheld). Note that Figure 6 depicts a subset of the (metric, matching score) correlations for each of the three algorithms. The shaded regions within Table 4 correspond to correlations in Figure 6. For the eight metrics, the correlation coefficients marked in bold (Table 4) are the highest. The bold items reaffirm that $minC$ has the highest correlation for all single-video factor metrics across all three algorithms.

Table 4. Correlation (%) of each of the four factors and eight metrics over the three algorithms (control+handheld)

Algs	Factors	M_k	Algorithm-dependent		Distribution-attribute	
			$lrpca_{best}$	$pittpatt_{best}$	μ	$\mu+\sigma$
PittPatt	Yaw	ΔY	-31	-10	-26	-26
		$maxY$	0	5	8	8
	Roll	ΔR	-8	-8	-15	-13
		$maxR$	-5	-8	-9	-8
	Size	ΔS	-3	-5	-17	-18
		$minS$	29	19	35	40
	Confidence	ΔC	-20	-19	-30	-27
		$minC$	**42**	**43**	**50**	**46**
PA	Yaw	ΔY	-30	-4	-27	-30
		$maxY$	3	7	8	9
	Roll	ΔR	-4	-2	-12	-10
		$maxR$	0	0	1	2
	Size	ΔS	5	1	-6	-8
		$minS$	16	12	23	30
	Confidence	ΔC	-20	-17	-30	-29
		$minC$	**41**	**40**	**49**	**49**
LRPCA	Yaw	ΔY	-38	-8	-35	-37
		$maxY$	7	12	13	13
	Roll	ΔR	-10	-4	-15	-14
		$maxR$	-5	-3	-6	-6
	Size	ΔS	-5	-2	-14	-16
		$minS$	18	8	19	25
	Confidence	ΔC	-21	-17	-30	-27
		$minC$	**54**	**51**	**62**	**60**

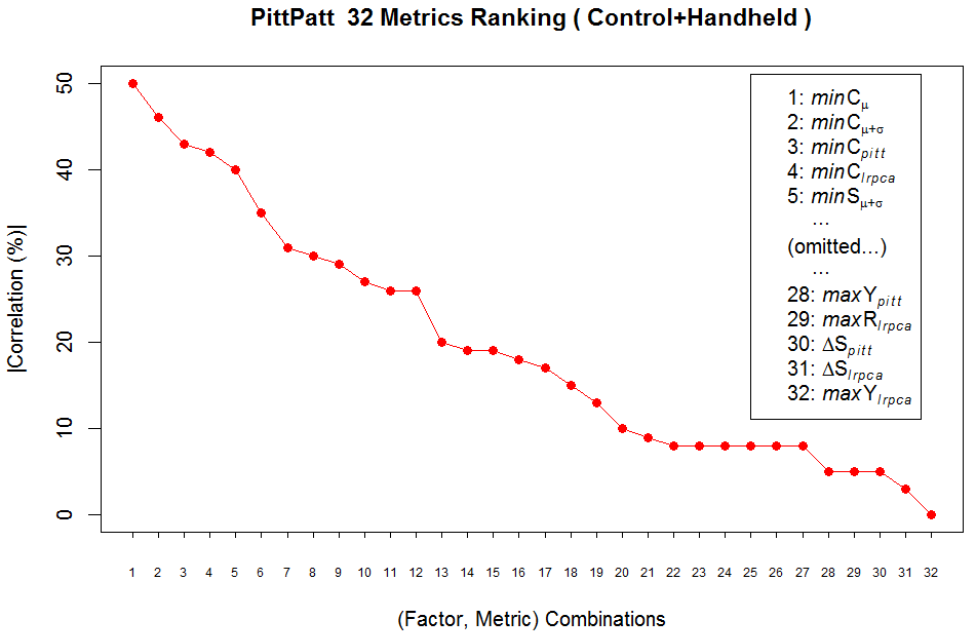

Figure 7. Ranking of the 32 (factor, metric) combinations based on |correlation| (%) of PittPatt's matching scores and metrics for the four factors (Yaw, Roll, Size, and Confidence)

To further assess the relative importance of the various (factor, metric) combinations, we sort and plot the |correlation coefficients| (%) based on Table 4; Figure 7 shows the ranks for all data for the PittPatt algorithm.

The x-axis is the 32 (factor, metric) combinations ordered by |correlation| (%) and the y-axis is the absolute correlation given as percentage (%). The legend shows only the top five and bottom five rankings. Figure 7 shows that the PittPatt matching score is most highly correlated (50 %) with $minC_{\mu}$, and is least correlated (0 %) with $maxY_{lrpca}$. For PittPatt, $minC$ is present in the top four rankings.

For all data (control+handheld), we investigate the rank number for the 32 (factor, metric) combinations for the PittPatt case only. For the comparative metrics, the results show that Δ is the best for Yaw and Roll, while the extreme (min) is the best for Size and Confidence.

To further highlight the ranking of the PittPatt case in Table 4, we extracted the best comparative metric for each of the four factors and reordered the rows by the row average and the columns by the column average—note that PittPatt's best value ("50") in Table 4 leads to the best ranking ("1") in Table 5.

Table 5. Ranking (1=best) of the chosen comparative metrics with the four single-video metrics for PittPatt (control+handheld)

Factors	M_k	μ	$\mu+\sigma$	$lrpca_{best}$	$pittpatt_{best}$	Average
Confidence	$minC$	1	2	4	3	2.5
Size	$minS$	6	5	9	14	8.5
Yaw	ΔY	11	11	7	20	12.3
Roll	ΔR	18	19	22	22	20.3
Average		9.0	9.3	10.5	14.8	

Table 5 indicates that Confidence/$minC$ has the highest correlation with algorithm performance (the average rank of 2.5) followed by Size/$minS$ (8.5), Yaw/ΔY (12.3), and Roll/ΔR (20.3). For single-video metrics, the distribution-based metric μ has the highest rank (9.0) and $\mu+\sigma$ the second highest (9.3). On the other hand, the algorithm-dependent metric $PittPatt_{best}$ has the lowest average rank (14.8), even for the PittPatt algorithm.

To assess the robustness of our conclusions, Figure 8 illustrates the ranked (factor, metric) combinations for all three algorithms.

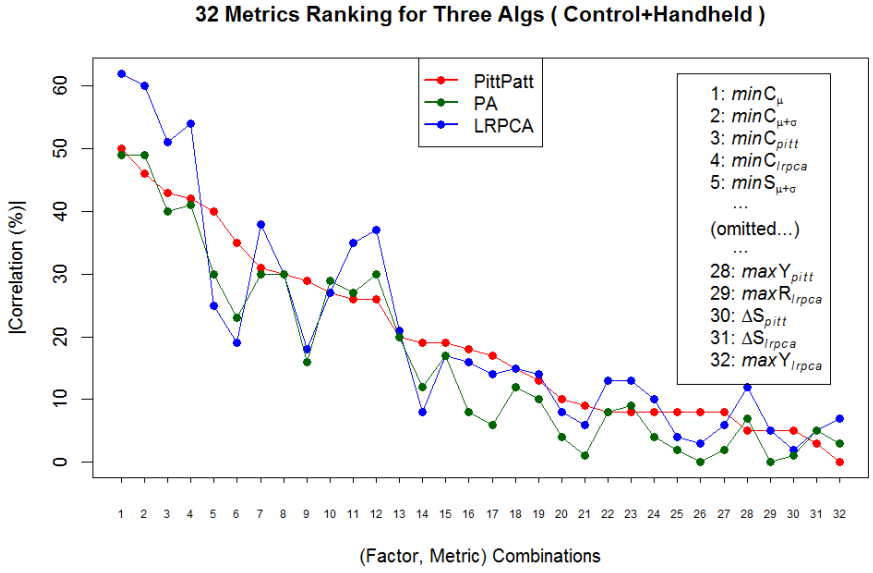

Figure 8. Ranking of the 32 (factor, metric) combinations for all three algorithms—ordered by the PittPatt results

The red dotted line is the ranking for PittPatt, green for PA, and blue for LRPCA. The (factor, metric) combinations are ordered by the PittPatt results. Figure 8 indicates that the best two metrics ($minC_{\mu}$ and $minC_{\mu+\sigma}$) are the best for all three algorithms. With obvious exceptions, the general trend of all three algorithms is similarly decreasing. Although algorithm-dependent metrics were yielded from the LRPCA or PittPatt algorithms, these algorithm matching scores still have a higher correlation with either $minC_{\mu}$ and $minC_{\mu+\sigma}$, which emanate from distribution-attribute type.

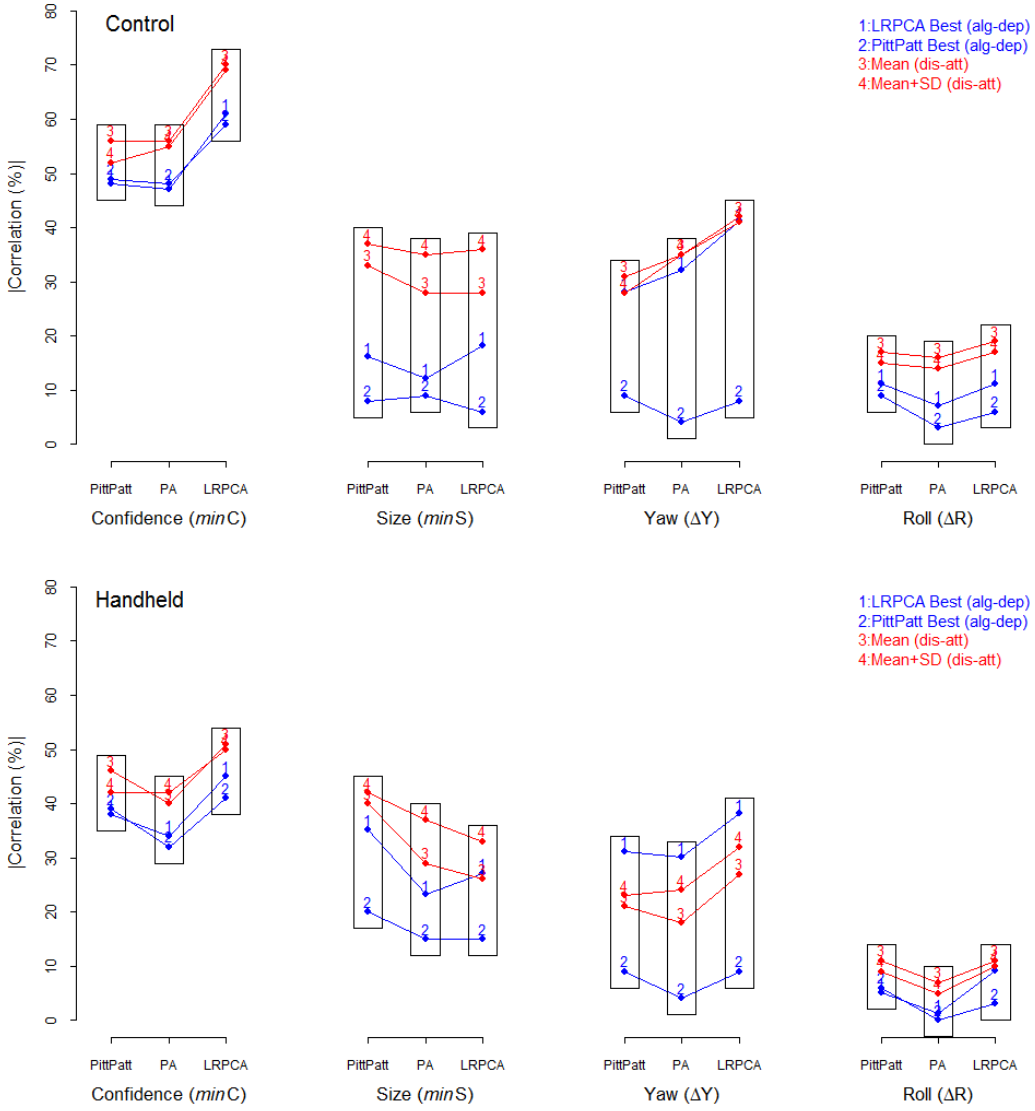

Figure 9. Comparison of the algorithm-dependent (alg-dep) and distribution-attribute (dis-att) approach for all three algorithms for each of the control and handheld; (top) Control, (bottom) Handheld

From the PittPatt results in Table 5, we see that the distribution-attribute metrics (μ and $\mu + \sigma$) are most highly correlated with the matching scores. We now reaffirm/extend these conclusions of robustness to the other algorithms (PA and LRPCA) for both the control-only and handheld-only data. We are primarily interested in the following two questions:

1) Which of the four metrics (1: $lrpca_{best}$, 2: $pittpatt_{best}$, 3: μ, and 4: $\mu + \sigma$) is most highly correlated with algorithm matching scores for the three algorithms and for each of the control and handheld data?
2) Which of the two approaches (algorithm-dependent vs. distribution-attribute) is most correlated with the matching scores for the three algorithms and for each of the control and handheld data?

To this end, we use the block plot shown in Figure 9; the top plot is based on the control data ordered by Confidence ($minC$), Size ($minS$), Yaw (ΔY), and Roll (ΔR); the botton plot is based on the handheld data. The four items within each block are the four metrics, with blue indicating algorithm-dependent and red indicating distribution-attribute.

For Figure 9 (top), the x-axis consists of the three algorithms (PittPatt, PA, and LRPCA) for each of the four factors. The y-axis is the |correlation| (%) of the four metrics with the matching scores.

Across all three algorithms over the control and handheld data (except the Yaw (ΔY) for the handheld case), overall, the distribution-attribute metrics (μ and $\mu + \sigma$) perform better than the algorithm-dependent metrics ($LRPCA_{best}$, $PittPatt_{best}$).

In summary, for video-based face recognition, the distribution-attribute approach can be more effective than the algorithm-dependent approach for measuring factor values. We also observe that the face detection confidence followed by face size may serve as a quality measure metric for predicting video-based face recognition performance.

5.2 Cross-Domain Environment and Sensor Analysis

In this section, we conduct a cross-domain analysis by including the Environment and Sensor factors. The cross-domain analysis here indicates examining the effects of cross-environment or cross-sensor on the performance of algorithms. This analysis addresses the following questions:

1) What is the ranking of algorithm performance across all (query and target) combinations of environments or sensor types?
2) Are the environment and sensor conclusions robustly true for all three algorithms (PittPatt, PA, LRPCA)?
3) Can we characterize algorithm performance based on environment or sensor type (e.g. indoor/ outdoor or control/handheld effects)?

Since the environment factor is confounded by different sensor models, we examine a cross-environment effect for control and handheld video data separately.

Figure 10 shows the results of the environment analysis for control (top) and handheld (bottom)—note that the control camera was used in all six environments, while other handheld cameras were used in only one or two environments, with obvious confounding implications. Considering query and target pairs, the PaSC video dataset yielded a total of 16 environment combinations (out of 36 possible).

The x-axis illustrates these16 combinations, and the y-axis is the VR at FAR=0.01. The red line is the results for PittPatt, green for PA, and blue for LRPCA—VR results are ordered by PittPatt's verification rate. For example, in the control case, the VR of the far-right pair {Phone, Paper} is 93 % verification rate for PittPatt, 57 % for PA, and 38 % for LRPCA.

For both control and handheld cases, in general, the performance follows a similar pattern for all three algorithms—the pair {Canopy, Canopy} has the lowest VR for PittPatt and LRPCA, and {Ball, Canopy} for PA. For the control case, the pair {Phone, Paper} has the highest performance for PittPatt (93 %), {Ball, Paper} for PA (65 %), and {Phone, Easel} for LRPCA (55 %). For the handheld case, {Phone, Easel} has the highest performance for both PittPatt (92 %) and LRPCA (49 %) and {Phone, Paper} for PA (62 %). For all three algorithms, the results from the control data have generally higher performance than the handheld data.

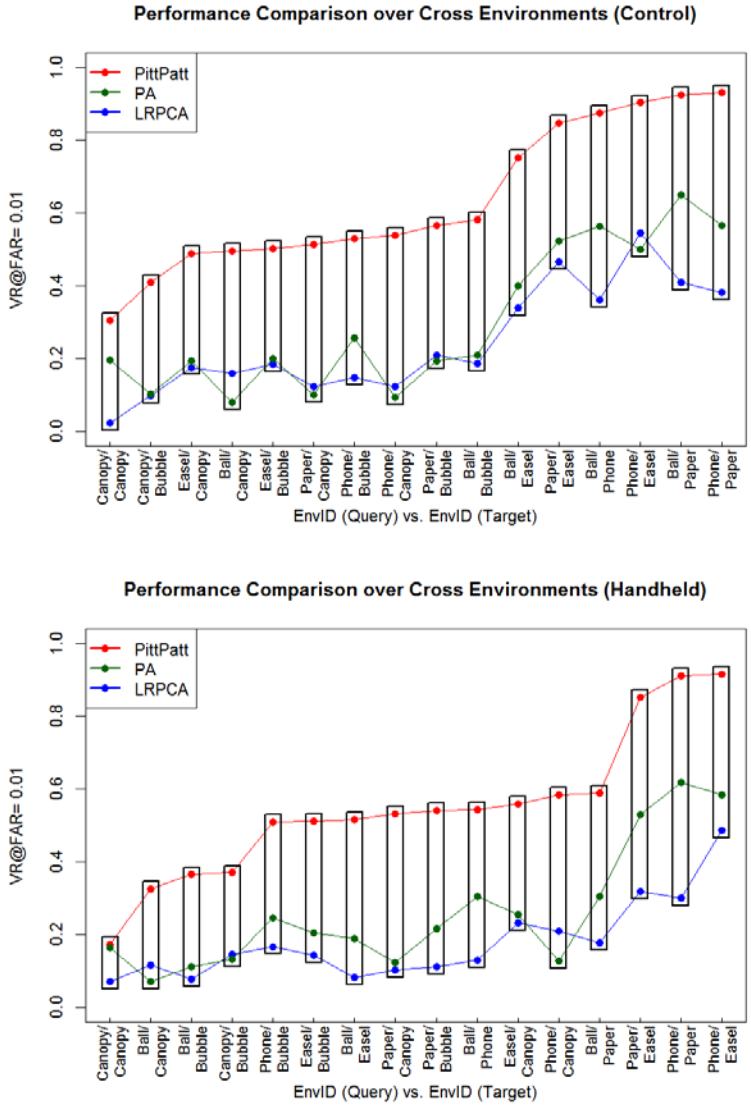

Figure 10. Performance ranking by environment combinations (16 pairs of query and target) for PittPatt, PA, and LRPCA — performance are ordered by PittPatt; (top) Control (bottom) Handheld

We also observe that all Canopy or Bubble pairs have a lower performance, while Ball, Phone, Paper, and Easel have a higher performance. This observation is robustly true for all three algorithms for both control and handheld data. Interestingly, both Canopy and Bubble scenes were taken outdoors while the others (Ball, Phone, Paper, and Easel) were taken indoors.

Next, we investigate how a cross-sensor type affects performance for all three algorithms.

Figure 11 shows the results of the six different sensor models. The x-axis represents all query and target combinations (13) across sensor models; note that the pair {C, C} is only from the control camera and the rest are pairs from the handheld camera.

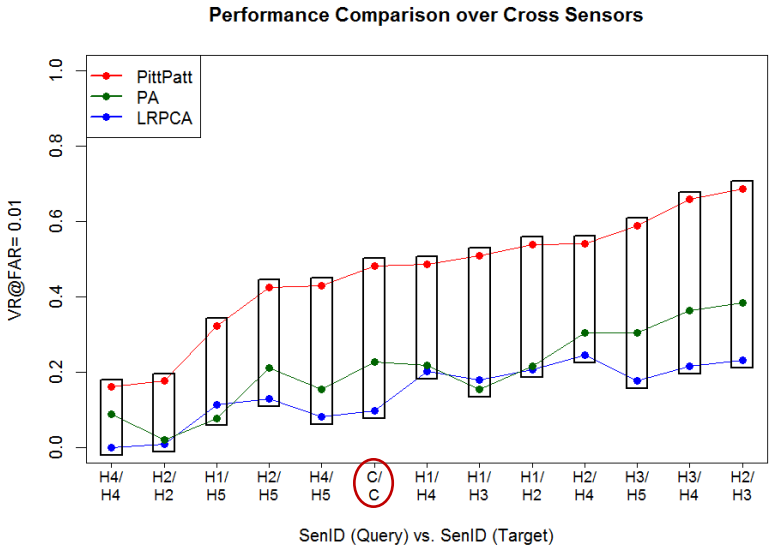

Figure 11. Performance ranking by the sensor (query and target) combinations (13 pairs) for all three algorithms—performance is ordered by PittPatt

The performance ranking for sensor pairs has a relatively similar pattern for all three algorithms. The results show that pair {H4, H4} has the lowest performance for both PittPatt (16 %) and LRPCA (0 %), and {H2, H2} for PA (2 %). On the other hand, {H2, H3} has the highest for both PittPatt (69 %) and PA (38 %), and {H2, H4} for LRPCA (25 %).

Although the pair {C, C} marked in circle is from the control camera, their verification rates are not best for all three algorithms—note that "C" is the only sensor that was used by all six different environments (see Table 1 and Table 2). This result indicates that the Environment and Sensor factors may have an interaction effect (due to environment and sensor confounding in the PaSC video dataset).

5.3 Subject ID, Gender and Race Analysis

In this section, we address the following questions for subjectID, Gender, and Race factors.

1) How do the three algorithms perform by subject?
2) Is algorithm performance robust over male vs. female?
3) Is algorithm performance robust over race?

To address the first question, Figure 12 provides the ranks of the matching scores by subject for all three algorithms for both datasets (control+handeld).

The x-axis is the subject ID of 265 individuals with the total number of videos for each person. For example, 6082/20 indicates that a total of 20 videos were taken from the subject ID 6082. The y-axis is the median value of the similarity (matching) scores out of the total number of videos per person—the scores for the three algorithms are normalized to the interval (0, 1), and a higher score implies a better similarity for all three algorithms. The red line represents the results for PittPatt, green for PA, and blue for LRPCA—265 subjects are ordered by the PittPatt's similarity scores.

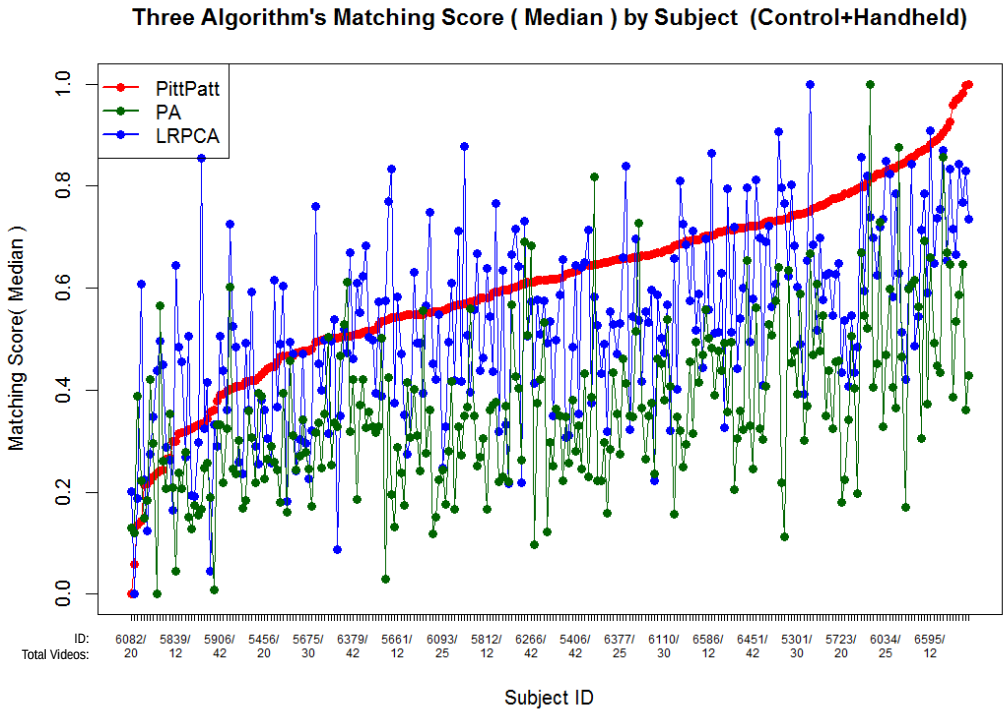

Figure 12. The three algorithms performance by 265 subjects (control+handheld); the x-axis is the subject ID with the total number of videos for each subject; the y-axis is the median value (out of a total videos) of the matching scores per subject.

Our results show that the pattern of algorithm performance by subject is markedly different amongst the three algorithms—the best subjects for PittPatt may not be the best subjects for PA and LRPCA. This conclusion is consistent across all (control+handheld), control-only, and handheld-only datasets.

Next, we investigate how gender affects algorithms performance. Out of 265 subjects, the PaSC video data contains 146 males and 119 females. Figure 13 shows algorithm performance by gender. The x-axis illustrates the three algorithms for control and handheld separately and the y-axis is the VR at a FAR=0.01. Across all six cases, we observe that males have higher verification rate than females.

Lastly, we examine the three algorithms performance by race. Out of 265 subjects, 214 people are Caucasians (White) and 33 people are Asian-Pacific (Asian). Due to a small number of subjects for Black-American (5), Hispanic (1) and Unknown (12), we omit these races in our evaluation. Figure 14 shows that, for both control and handheld data, Asians have higher verification rate than Whites, and the conclusion is robustly consistent across all three algorithms. For video-based face recognition, we thus observe that Asians are easier than Caucasians for verification.

Though independence issues are a consideration for the gender and race analysis, the sample sizes are markedly large, which yield negligibly small confidence limits.

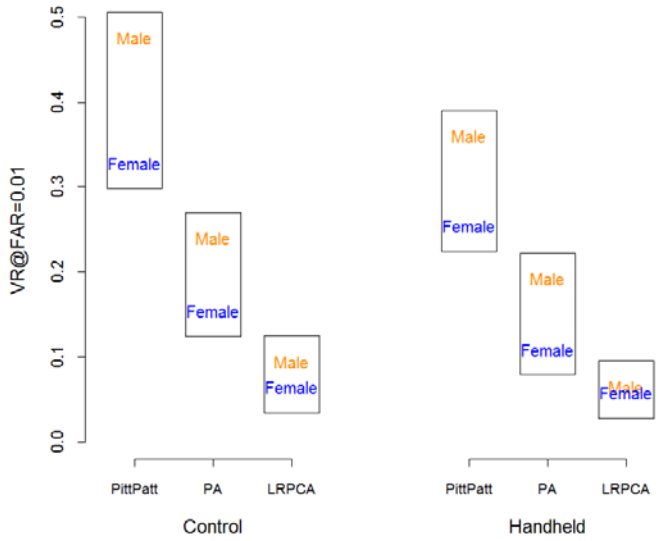

Figure 13. The three algorithms performance by gender

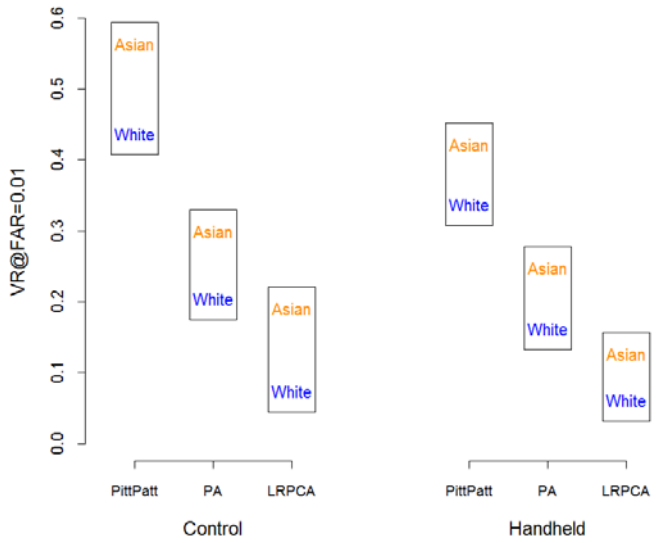

Figure 14. The three algorithms performance by race

5.4 Sensitivity Analysis for the Nine Factors

This section discusses the ranked list of important factors on algorithm performance. We investigate each factor effect and compare its relative effect on three algorithm's performance. Figure 15 shows main effects plots for the nine factors for the combined control and handheld video data; the first plot (a) is the results for PittPatt, the second plot (b) for PA, and the last plot (c) for LRPCA.

Note that the results from control-only and handheld-only data are relatively similar to the results from all (control+handheld) data; we thus illustrate the main effect plot for all (control+handheld) only.

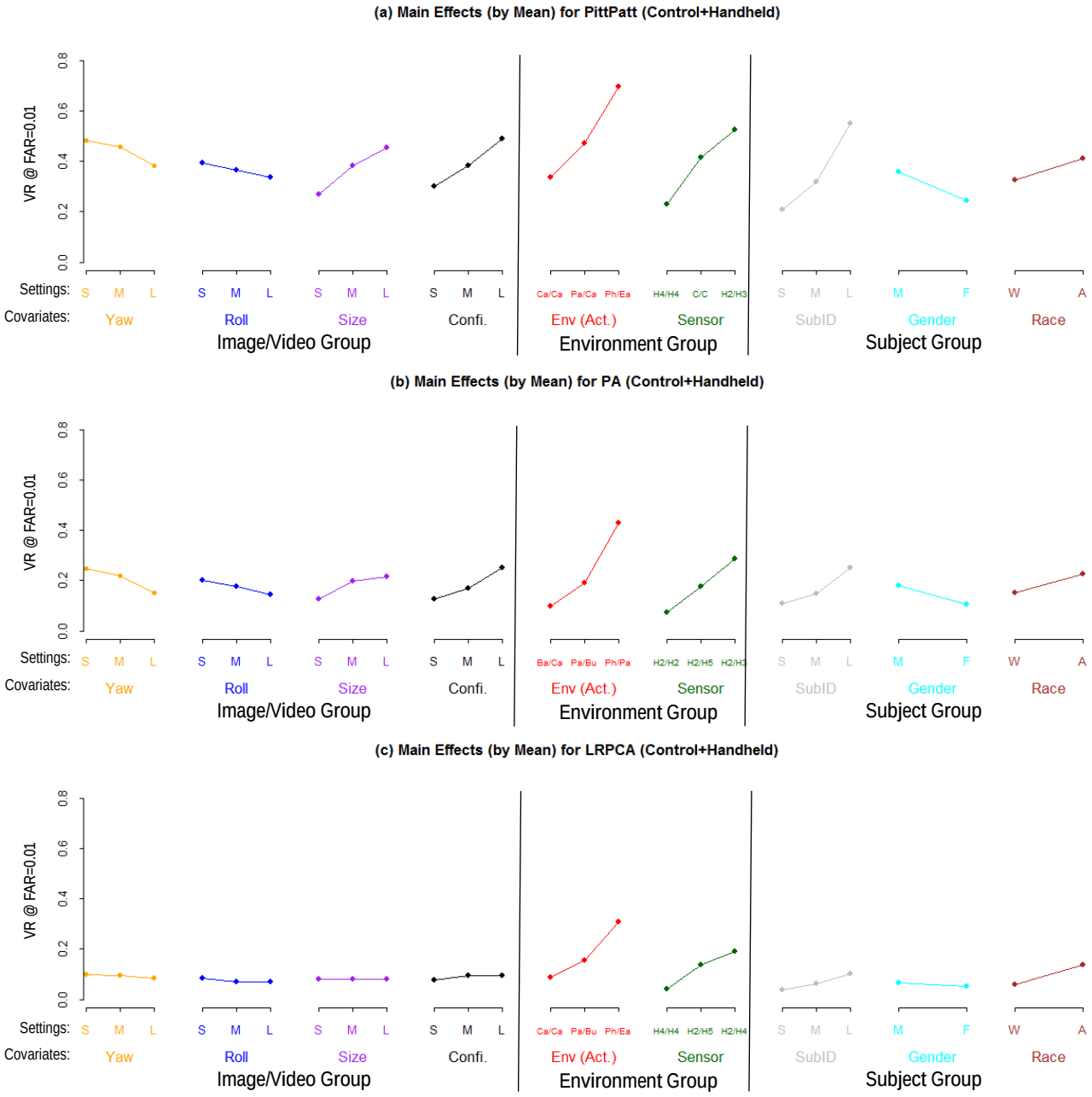

Figure 15. Summary of the nine factors effect for (a) PittPatt, (b) PA, and (c) LRPCA from all video data

In Figure 15 (a), the x-axis lists the factor name and its (selected) discretized level settings, and the y-axis is the response (VR at FAR=0.01). A steeper sloped line (large magnitude) indicates that a factor has a greater effect on the performance while a flatter line indicates that the factor has a lesser effect.

Based on the results of Section 5.1, the four factors within the image group are computed from the single-video factor metric "mean" with the comparative metric Δ for Yaw and Roll, min for Size and Confidence. The choice of the level settings was done as follows: for each of the four factors, we marginally divided the settings into three

levels: S (small), M (middle), L (large). For the two factors from the environment group, we select VRs of the lowest, medium, and highest out of all combinations. In the PittPatt case, for the Environment factor, the lowest pair is {Canopy, Canopy}, the medium {Paper, Canopy}, and the highest {Phone, Easel}. For Sensor, the lowest of the pair is {H4, H4}, the median {C, C}, and the highest {H2, H3}—see details in Section 5.2. Out of three factors within the subject group, the SubjectID settings were divided into the three levels (S: Small, M: Middle, L: Large) based on the ordered individual performance. The gender factor has two levels (M: male, F: female), and the race with the two levels (W: White, A: Asian)—see details in Section 5.3.

Within the image group, for PittPatt, Confidence and Size have the highest effect on algorithm performance followed by Yaw. For PA, these four factors have less effect on performance and the most important factor is Confidence and Size. For LRPCA, these four factors have little effect on algorithm performance. Within the environment group, the Environment with Activity factor has slightly higher effect than Sensor. For the subject group, SubjectID (individual) has the highest impact on performance for PittPatt but a lesser effect for PA and LRPCA. For each of the three groups, the environment group has the highest effect on performance across all three algorithms. These conclusions are robustly consistent over all three algorithms. Though independence issues are a consideration, the pair (query and target) sample sizes are markedly large, which yield negligibly small confidence limits and corresponding statistical significance.

For all (control+handheld) videos, the ranked list of the relative importance of the nine factors for each of the three algorithms is as follows.

For PittPatt:

1) Environment with Activity
2) SubjectID
3) Sensor
4) Confidence, Size, Yaw
5) Gender, Race, Roll

For PA:

1) Environment with Activity
2) Sensor
3) SubjectID
4) Confidence, Size, Yaw
5) Gender, Race, Roll

For LRPCA:

1) Environment with Activity
2) Sensor
3) Race, SubjectID, Confidence, Yaw, Size, Gender, Roll

In summary, the ranked lists are relatively similar across all three algorithms. Out of the nine factors, Environment with Activity is the most important and {Gender, Race, and Roll} are the least important factors—this conclusion is robustly true across all three algorithms for both control and handheld.

6. Discussions and Conclusions

We presented an analysis method to examine factor effects on face recognition algorithms in a video. Using the PaSC video dataset, nine factors from three groups (image/video, environment, and subject) were investigated to examine their impacts on the performance of three algorithms (PittPatt, PA, and LRPCA). We also introduced and studied four single-video and two comparative factor metrics for characterizing face recognition algorithms in a video.

For the comparative metrics, and for the four factors (Yaw, Roll, Size, and Confidence) within the image group, we found that the "extremum" approach performed better for Confidence ($minC$) and Size ($minS$), while the "difference" approach performed better for Yaw (ΔY) and Roll (ΔR). For single-video metrics, the distribution-attribute metrics (μ and $\mu + \sigma$) performed better than the algorithm-dependent (frame-to-frame) metrics—thus,

the distributional approach is generally more effective to quantify factor values for video-based face recognition. These conclusions were robustly consistent over all three algorithms and for each of control and handheld videos.

We also observed that the face detection confidence followed by face size can potentially serve as a quality measure metric for predicting face recognition performance in video.

We next conducted a cross-domain analysis for the two factors (Environment and Sensor) within the environment group. For this environment factor, we reaffirmed that videos taken outdoors are indeed more challenging than indoors for recognizing a person using the face—this conclusion agreed with a previous study on the FRVT2006 dataset done by Beverage et al. [4]. For the Sensor factor, the pair of the control video data performed no better than other handheld data—thus, we found that the environment and sensor factors have an interaction effect due to these two factors being confounded in the PaSC dataset.

For the subject group, the trend of algorithm performance by the SubjectID factor is markedly different over the three algorithms. For gender, we also found that male had a higher verification rate than female, and for race, Asian had a higher verification rate than White. These gender and race conclusions were robustly consistent across all three algorithms, and for each of the control and handheld datasets.

In summary, over the nine factors, Environment with person's activity had the most effect on algorithm performance in a video, followed by Sensor and SubjectID—these conclusions were true for each of the three algorithms, and for both control and handheld. We thus conclude that scene/action actually matters for face (human) recognition—this implies reinforcing the importance of isolating a subject from the background, or characterizing a subject's action.

Acknowledgement

The authors would like to thank Dana Udwin for her assistance in algorithm evaluations and her hard work during the SURF (Summer Undergraduate Research Fellowship) program at NIST, 2013.

Disclaimer

The identification of any commercial product or trade name does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

References

[1] S. K. Zhou and R. Chellappa, "Beyond one still image: Face recognition from multiple still images or a video sequence," *Face Processing: Advanced Modeling and Methods*, pp. 547–567, 2005.

[2] G. Givens, J. R. Beveridge, B. A. Draper, P. Grother, and P. J. Phillips, "How features of the human face affect recognition: a statistical comparison of three face recognition algorithms," in *Computer Vision and Pattern Recognition (CVPR)*, 2004, vol. 2, pp. II–381.

[3] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. J. Phillips, "A meta-analysis of face recognition covariates," in *Biometrics: Theory, Applications, and Systems (BTAS)*, 2009, pp. 1–8.

[4] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, and Y. M. Lui, "Focus on quality, predicting FRVT 2006 performance," in *8th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2009.

[5] P. J. Phillips, J. R. Beveridge, D. S. Bolme, B. A. Draper, G. H. Given, Y. M. Lui, S. Cheng, M. N. Teli, and H. Zhang, "On the existence of face quality measures," in *Biometrics: Theory, Applications and Systems (BTAS).* 2013.

[6] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. S. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer, "An introduction to the good, the bad, & the ugly face recognition challenge problem," in *Automatic Face & Gesture Recognition and Workshops (FG).*, 2011, pp. 346–353.

[7] Pittsburgh Pattern Recognition, "PittPatt SDK v5.2.2 Documentation: Detection User's Guide." .

[8] O. Yamaguchi, K. Fukui, and K. –. Maeda, "Face recognition using temporal image sequence," in *Automatic Face and Gesture Recognition (FG)*, 1998, pp. 318–323.

[9] J. R. Beveridge, B. A. Draper, J.-M. Chang, M. Kirby, H. Kley, and C. Peterson, "Principal angles separate subject illumination spaces in YDB and CMU-PIE," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 351–363, 2009.

[10] P. J. Grother, G. W. Quinn, and P. J. Phillips, "Report on the evaluation of 2d still-image face recognition algorithms," *NIST Interagency Rep*, no. 7709, 2010.

[11] J. R. Beveridge, P. J. Phillips, D. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, and K. W. Bowyer, "The Challenge of Face Recognition from Digital Point-and-Shoot Cameras," *Biometrics: Theory, Applications, and Systems (BTAS)*, 2013.

[12] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 Large-Scale Experimental Results," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 831–846, 2010.

[13] R. B. Vaughn Jr, R. Henning, and A. Siraj, "Information assurance measures and metrics-state of practice and proposed taxonomy," in *Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003.*

[14] Y. Lee, J. J. Filliben, R. J. Micheals, and P. Jonathon Phillips, "Sensitivity analysis for biometric systems: A methodology based on orthogonal experiment designs," *Computer Vision and Image Understanding*, vol. 117, pp. 532–550, 2013.