**NISTIR 8051**

# It's About the Face Impostor Distribution

P. Jonathon Phillips
Amy N. Yates
Geof H. Givens
J. Ross Beveridge

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8051

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

**NISTIR 8051**

# It's About the Face Impostor Distribution

P. Jonathon Phillips
Amy N. Yates
*Information Access Division*
*Information Technology Laboratory*

Geof H. Givens
J. Ross Beveridge
*Colorado State Unviersity*
*Fort Collins, CO*

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8051

April 2015


U.S. Department of Commerce
*Penny Pritzker, Secretary*

National Institute of Standards and Technology
*Willie May, Acting Under Secretary of Commerce for Standards and Technology and Acting Director*

# It's About the Face Impostor Distribution

P. Jonathon Phillips
NIST
100 Bureau Dr
Gaithersburg, MD 20899
jonathon@nist.gov

Geof Givens
Dept. Statistics
Colorado State U.
Fort Collins, CO 80523
geof@stat.colostate.edu

Amy N. Yates
NIST
100 Bureau Dr
Gaithersburg, MD 20899
amy.yates@nist.gov

J. Ross Beveridge
Dept. Computer Science
Colorado State U.
Fort Collins, CO 80523
ross@cs.colostate.edu

## Abstract

*We studied the effects of factors on the false accept rate (FAR) for three modern video face recognition algorithms. We examined the effects of environment (location), video-(imagery-) based, and demographic factors. The study is performed on the handheld video in the Point and Shoot Face Recognition Challenge (PaSC), which consists of 1401 handheld videos of 265 subjects. The results of our analysis are consistent across the three algorithms. Our analysis shows that FAR can significantly vary. Surprisingly, for environment and video-based factors, there was a clear relationship between verification rate (VR) and FAR. An increase (resp. decrease) in the FAR results in an increase (resp. decrease) in the VR. We looked at the shape of the marginal impostor distributions for each level of a factor. In most cases these impostor distributions for a given algorithm moved according to a simple affine transform, translation and scaling, when moving between factor levels.*

## 1. Introduction

Faces are highly variable. A face can be smiling or angry; a face can be in bright sunlight or in a poorly lit room; a face can be viewed in a still image or a YouTube video. Attributes of people may also matter. Are they Caucasian or East Asian, old or young, or healthy or ill? All of these changes can combine to make face recognition easier or harder. The challenge is knowing which of these factors most influence performance. By identifying the factors with the greatest impact, it is possible to focus algorithm research and understand how algorithms will respond when being used. The research community currently tends to focus attention on three factors: pose, expression and illumination [6], and these are clearly important. Recently, however, environment, the combination of location and sensor, has been shown to strongly effect verification rate (VR) [8].

The analysis presented here quantifies how environmental, video-based and demographic factors effect the impostor distribution, i.e. the likelihood of falsely matching pairs of faces of different people. Studying the impact of factors on the impostor distribution, and thus the false accept rate (FAR), allows us to ask a series of questions. Do changes in factors effect FAR? Is there a relationship between VR and FAR? Are some environments (locations) easier than others? Are larger faces easier to recognize? How do changes in a factor effect the impostor distribution? Do these effects alter the shape of the tail of the impostor distribution? All of these questions are addressed in the results presented below.

The key contributions of this work are:

1 The first comprehensive study of how factors effect the impostor distribution and FAR for video face recognition algorithms. We examine environment, video-based (imagery-based), and demographic factors.

2 The FAR varies significantly. For the environment factor, the FAR varies from 0.01 to 0.43 for one algorithm, and from 0.03 to 0.27 and from 0.04 to 0.24 for the two other algorithms.

3 For environment and video-based factors, there was a clear relationship between verification rate (VR) and FAR. An increase (resp. decrease) in the FAR results in an increase (resp. decrease) in the VR.

4 A study of how changes in the environment and video-based factors effect the impostor distribution.

The video-based factors are computed by an algorithm. Since these factors are computational, the results in model key conditions in real-world scenarios and the results in this paper could be tested in these scenarios.

In prior work, how factors, covariates, effect the match distribution and consequently verification rates has been studied, and a mature protocol is in place for analyzing face recognition performance based upon match-pairs [1],[5]. In contrast, the impostor distribution has been less studied. In the Good, the Bad and the Ugly (GBU) Challenge problem the impostor distribution was examined and shown to be relatively stable across the three partitions of varying difficulty [12]; thus it was not emphasized. However, more recently O'Toole et al. [11] looked at the effect of race and gender on the impostor distribution. They found that performance changes when the impostor distribution is restricted to people of the same gender or race. Sgori et al. [13] created a version of the GBU based on the impostor distribution.

Adding to the evidence that the impostor distribution does matter, in this paper we present a detailed analysis of impostor distributions for the video portion of the Point and Shoot Face Recognition Challenge (PaSC) [2]. The PaSC contains video taken with handheld video cameras that are typical of those in cell phones. The PaSC is a designed data set which systematically varied imaging factors including camera, location and subject action. Also, included in our analysis are subject and imagery factors. The imagery factors are based on distributional analysis of a video sequence [8]. This allows us to compute measures of yaw and face size for a video that are predictive of performance. To the best of our knowledge, this is the first comprehensive study of factors that effect the impostor distribution for unconstrained face recognition.

## 2. PaSC Challenge and Data Set

The analysis presented here is carried out for three algorithms applied to the handheld video portion of the Point-and-Shoot Challenge (PaSC) [2]. Section 2.1 summarizes the data including details about the video data. Section 2.2 summarizes the evaluation protocol and algorithms.

### 2.1. Video Data

The PaSC handheld video data was collected at the University of Notre Dame over seven weeks in the Spring semester 2011. In a given week, all videos were collected at the same location. The videos show people carrying out tasks rather than looking into a camera. Collection was carried out according to a plan - a script - in which generally a person entered a scene, approached some designated spot, carried out an action, and then left the scene. The videos typically begin as the person is moving into the scene and terminate as the person is leaving. Video length ranges roughly between 50 and 400 frames with most videos containing between 200 and 250 frames. There are a total of 1401 videos of 265 different people. Each person appears in between 4 and 7 videos. Videos were acquired at 6 different locations using one of five different cameras at resolutions ranging between 640×480 to 1280×720. Thus, each video represents one person at a specific location captured with one handheld video camera. Table 1 summarizes the camera/location/action combinations[1].

Table 1. Location, camera and action combinations. The abbreviations for the environment is in the right column.

| Sensor | Location | Action | Abbrev. |
|---|---|---|---|
| Flip Mino F360B | canopy | golf swing | Ca |
| Kodak Zi8 | canopy | bag toss | Ca |
| Samsung M. CAM | office | pickup newspaper | Pa |
| Sanyo Xacti | lab 1 | write on easel | Ea |
| Sanyo Xacti | lawn | blow bubbles | Bu |
| Nexus Phone | hallway | ball toss | Ba |
| Kodak Zi8 | lab 2 | pickup phone | Ph |

In the findings below, the influence that location and action combinations exert over performance is strong, and the abbreviations introduced in Table 1 will be used when reporting results. Therefore here, briefly, is a bit more information about each. The **canopy** (Ca) was a white pop-up material structure setup outside in bad weather. Two actions were carried out on different days. The first was swinging a golf club and the second tossing a bean bag. The **office** (Pa) was a large well lit room where a subject picked up and looked at a newspaper. In **Lab 1** (Ea) each subject wrote on a large floor standing easel set out in a large open lab space. The **lawn** (Bu) was an open grassy area in a plaza with bright sun. Subjects approached a table and blew bubbles. The **hallway** (Ba) was an interior space of an older building with relatively dark stone walls where subjects threw a toy basketball. In **lab 2** (Ph) a subject picked up a phone in a relatively cluttered lab area.

Figure 1 shows four zoomed-in clips from four different videos. The upper left clip is from the **office**. The upper right is from the **canopy**. The lower left is from **lab 2**, and the lower right is from the **lawn**. These frames are characteristic in several respects, for example suggesting the range of lighting conditions and also the fact that in general subjects are not attending to the camera.

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

Figure 1. Clips of two people sampled from four PaSC handheld videos. All four videos were taken at different locations: two outdoors and two indoors.

### 2.2. Algorithms

Our analysis is performed on three of the top performers in the Face and Gesture 2015 Person Recognition Evaluation [3]. The algorithms were developed independently by three groups. This independence provides evidence that our conclusion will generalize to algorithms not included in this study. Several other conditions were adopted to make sure the results were not tuned to the PaSC data set. First, the algorithms were not trained on subjects in the PaSC challenge. Second, the algorithms were not trained on imagery from locations that are included in the PaSC challenge. Third, cohort or gallery normalization using the PaSC imagery was not allowed.

Face detection and associated eye coordinates estimated by the Pittsburgh Pattern Recognition (PittPatt) face recognition SDK 5.2.2 were made available to algorithms used in this study. The three algorithms represent relatively distinct approaches as summarized below.

The Chinese Academy of Science (**CAS**) algorithm uses two convolutional neural networks, one for larger and one for smaller faces [7]. Network features are pooled and fed to three kernel linear discriminant analysis based-algorithms operating over sets of video frames. Similarity is the weighted sum of the cosine angle between query and target videos feature vectors.

The University of Ljubljana (**Ljub**) algorithm combines four feature types with a probabilistic principal component analysis [15]. The four feature types are Gabor wavelets, local binary patterns, local phase quantization histograms, and pixel intensity. Fixed sized templates are then generated for each video and finally compared using a linear logistic regression weighting scheme.

The Stevens Institute of Technology (**SIT**) algorithm combines scale-invariant feature transform (SIFT) features with a probabilistic modeling procedures and principal component analysis based dimensionality reduction process [9], [10]. The probabilistic modeling is realized through a mixture of Gaussians. Fixed sized templates result, one template per video, and these are compared using a joint Bayesian classifier.

## 3. Methodology for Analysis

Video meta-data divides into three basic categories. First, environment-pair factors arising from the combination of locations, sensors, and actions summarized above in Table 1. Second, video-based factors such as the size of the face as it appears in the videos or the degree the face image is frontal. Third, there is demographic information about the person in the video, in particular gender and race. The methodology set forth below allows us to quantify how changes in factors associated with this meta-data influence the likelihood of generating a false match.

### 3.1. Measuring Performance

Performance is measured for a verification task. In a verification task, two faces are presented to an algorithm, and the algorithms responses with a measure of the degree of similarity of the two faces. In this paper all faces are in videos. Formally, algorithm $A$ produces a similarity score $s_A(x, y)$ between two faces in videos $x$ and $y$. The two videos are referred to as face-pair $(x, y)$. We assume that there is one prominent face in each video. A larger similarity score indicates a higher likelihood that the faces in the image are the same.

Performance is computed over a set of face-pairs $F$ with the face-pairs divided into two sets: match-pairs $M$ and impostor-pairs $I$. A face-pair is a match-pair if the videos are of the same person and an impostor-pair otherwise. A verification decision is made by thresholding a similarity score and declaring the faces are the same if $s_A(x, y) \geq \tau$ and different otherwise. Note the similarity scores $s_A(x, y)$ for an Algorithm $A$ are contained in the similarity matrix produced by algorithm $A$. The verification rate (VR) for an algorithm is the fraction of match-pairs that are correctly declared to be the same person at a set threshold $\tau$; formally,

$$\mathrm{VR}(s_A(F), \tau) = \frac{\#\{s_A(x, y) \geq \tau \text{ and } (x, y) \in M\}}{\#\{(x, y) \in M\}} \quad (1)$$

The false accept rate (FAR) is the fraction of impostor-pairs

that are incorrectly declared the same person; formally,

$$\mathrm{FAR}(s_A(F), \tau) = \frac{\#\{s_A(x,y) \geq \tau \text{ and } (x,y) \in I\}}{\#\{(x,y) \in I\}} \quad (2)$$

To quantify how changing a factor alters performance these equations need to be extended. Let us illustrate using gender as an example. The subject in one video must be either male of female: a two-level factor. Therefore, the first step is to create a gender factor with three levels: both-female, both-male, or female-male. Note female-male only arises for impostor-pairs. The next step is to establish a global threshold $\tau_g$. In this paper, $\tau_g$ is usually chosen so that $\mathrm{FAR}(s_A(F), \tau_g) = 0.1$. [2]

Now, given a global threshold $\tau_g$ and factor levels $E_i$, the marginal VR and FAR broken out by factor levels are:

$$\mathrm{VR}(s_A(E_i), \tau_g) = \frac{\#\{s_A(x,y) \geq \tau_g \text{ and } (x,y) \in E_i \cap M\}}{\#\{(x,y) \in E_i \cap M\}}, \quad (3)$$

and

$$\mathrm{FAR}(s_A(E_i), \tau_g) = \frac{\#\{s_A(x,y) \geq \tau_g \text{ and } (x,y) \in E_i \cap I\}}{\#\{(x,y) \in E_i \cap I\}}. \quad (4)$$

Thus, for the gender example, there are separate marginal VR and FAR values for $E_{\text{both-female}}$, $E_{\text{both-male}}$, and $E_{\text{female-male}}$. These marginal rates quantify how VR and FAR change when changing between factor levels.

### 3.2. Environment, Video & Demographic Factors

The environment-factor arises because of the data acquisition plan and specifically the seven distinct location, sensor, and action combinations summarized Table 1. Also, since the environment factor must describe pairs of videos, there are a total of 22 pairs. In 15 environment-pairs, the videos are from different environments and acquired on different weeks; in one pair, the videos are from the same environment (canopy) and acquired on different weeks. For 6 pairings the videos are from the same environment and collected in the same week; these only include impostor pairs, i.e. pairs of videos of different people.

Video-factors derive from single image properties, for example the face size measured as the number of pixels between the eyes. Another is yaw, the degree to which the face is turned to the side. Estimates of both are available from the PittPatt SDK 5.2.2 software, as is one more factor, the confidence of the algorithm that a true face has been found. All three of these are of interest. To map these factors computed on a per frame basis to a single value characterizing a whole video we adopt the distributional method of Lee et al. [8]. To explain, if $g_I(x)$ is the factor for image $x$, then for a set of video frames $\{f_1, \ldots, f_n\}$ and factor values $\{g_I(f_1), \ldots, g_I(f_n)\}$ the video factor is $g_V(x) = \mathrm{mean}\{g_I(f_1), \ldots, g_I(f_n)\}$. The prior work of Lee et al. [8] showed this extension to video was useful.

The video-factors are extended to pairs of videos as follows. For yaw, the factor $h_{yaw}(x,y) = |g_V(x) - g_V(y)|$ where $g_V(x)$ and $g_V(y)$ are the yaw factors for videos $x$ and $y$. The yaw factor is larger between pairs of videos when they show, on average, faces from less similar viewpoints. The extension to pairs for face size and confidence take the smaller value from the pair: $h(x,y) = \min\{g_V(x), g_V(y)\}$. The assumption is the smaller of the two values is the dominant predictor of recognition difficulty. The real-valued factors are converted to levels by ordering video-pairs from smallest to largest factor value and then dividing them into $n$ equal sized bins. The result is $n$ levels ranging from smallest to largest factor value.

The demographic-factors, specifically gender and race, are encoded as already suggested above in the example from Section 3.1. For gender the levels are female-male (F/M), male-male (M/M) and female-female (F/F). For race they are Caucasian-Asian (C/A), Caucasian-Caucasian (C/C) and Asian-Asian (A/A).

## 4. Results for Environment-Pair Factors

It is well known that environment significantly effects algorithm performance. The design of the PaSC data set enabled us to characterize the impact of environment on performance. Previous studies have investigated the effect of environment on verification rates [1], [8]. We proceed by examining the effect of environment on the FAR and then look at the relationship between FAR and VR.

Since comparisons are between two videos, we look at performance for environment-pairs. For the three algorithms in our study, we computed the FAR for the 22 environment-pairs as described in Section 3.1. Figure 2 summarizes how environment-factors effect FAR for the three algorithms. Along the horizontal axis the 22 pairs of environments described in Section 3.2 are enumerated. The vertical axis shows the marginal FAR values, eq. 4, using a $\tau_g$ that corresponds to a global FAR = 0.10. The environment-pairs are ordered by marginal FAR averaged over the three algorithms. All environment-pairs to the left of vertical line, from pairs Ba-Ca to CaDW-CaDW, are cross-week pairs: CaDW signifies canopy videos taken in different weeks. All pairs to the right consist of video-pairs taken in the same week.

The principal finding is that environment exerts a dramatic influence over the impostor distribution and hence the marginal FAR. Algorithm **Ljub** has the greatest range in FAR from 0.01 to 0.43, and algorithm **CAS** has the smallest range from from 0.04 to 0.24. For cross-week environment-

[2] FAR=0.01 is the standard for PaSC when reporting VR. However, our goal is to quantify how factors alter FAR and VR; this shift in threshold better leverages the available data.

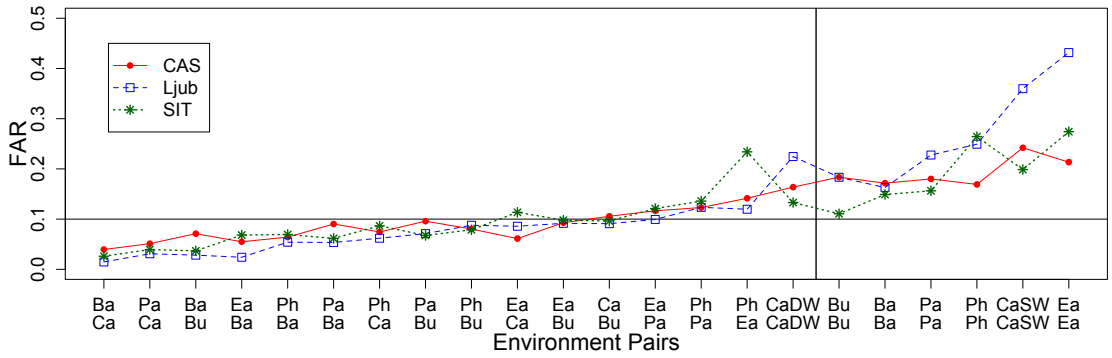


Figure 2. FAR of each environment-pair for each algorithm–ordered by mean FAR over all algorithms. The horizontal line corresponds to the global FAR = 0.10. The vertical line between pairs CaDW-CaDW and Bu-Bu separates the pairs into cross-week (left) and same week. The two canopy events were separated different-week and same-week pairs.

pairs, the range for **Ljub** is 0.01 to 0.22, **SIT** is 0.03 to 0.24, and **CAS** is 0.04 to 0.16. The FAR for the three algorithms varies by a factor of 22, 8, and 4 respectively. Prior work has already suggested the importance of environment [1], [8], this is the first clear evidence of how significantly it effects the impostor distribution.

A related finding is the importance of the cross-week versus same-week distinction. The mean cross-week marginal FAR averaged over the algorithms was 0.09 compared to 0.22 for cross-week pairs. A recent related result on still face image by Sgori et al. [13] also showed higher FAR values for same day image-pairs compared to different day image-pairs. One important conclusion is that the presence of impostor pairs in a data set taken at the same time biases upward the expected FAR for the data set as a whole.

### 4.1. Do VR and FAR Track Together?

We will now look at the relationship between the environment-pair FARs and VRs for the cross-week pairs. Scatterplots in Figure 3 relate marginal VR to marginal FAR, eqs. 3 and 4, for the 16 cross-week environment pairs. The horizontal axis is the FAR on a log-scale, and the vertical axis is the VR on a linear scale. The points represent environment-pairs, and the line is a linear regressor. For all three algorithms, the regression line suggests a linear relationship between $\log(\mathrm{FAR})$ and VR. In other words, an environment-pair that has a higher marginal VR will likely have a higher marginal FAR. Unfortunately, this linear relationship suggests that finding an environment-pair that is easier than others is unlikely. We say an environment-pair is easier if it has both a higher VR and a lower FAR than other pairs.

## 5. Results for Video-Based Factors

The impact of image- and video-based factors on verification rates have been extensively studied; however, their impact on the FAR has not been examined. We first look at the relationship between FAR and VR for three video-based factors and then investigate if there is an interaction between environment-pairs and the video-based factors.

Figure 4 shows the trade-off between FAR and VR for face size. The procedure described in Section 3.2 for creating factor levels through sorting and binning was used to create 10 face size factor levels: smallest faces to largest faces. Each point in Figure 4 is plotted according to the average marginal VR and FAR for all those video-pairs at one face size level. A trend similar to that seen for environment factors is evident, changes in face size associated with higher marginal VR correlate with higher marginal FAR. There is a similar relationship for yaw and face size; see Figure 9 in Supplemental Material for the corresponding scatterplots.

Figure 5 highlights possible interactions between environment and video factors for Algorithm **Ljub**. Like the scatterplots in Figure 3, each point corresponds to an environment-pair. Unlike in Figure 3, in Figure 5 circle size varies and is proportional the mean video factor for an environment-pair. For the yaw-factor, all the circles are about the same size, which means that yaw does not interact with the environment-pair. In contrast, a clear interaction effect between environment and face size is evident: environment-pairs with smaller VR and FAR tend to have small circle sizes and hence smaller mean face sizes. Figure 5 also suggests some interaction between environment and face confidence.

This analysis was repeated for Algorithms **SIT** and **CAS**, and the conclusions were the same. A complete set of plots for this analysis are in Figure 10 in the Supplemental Mate-

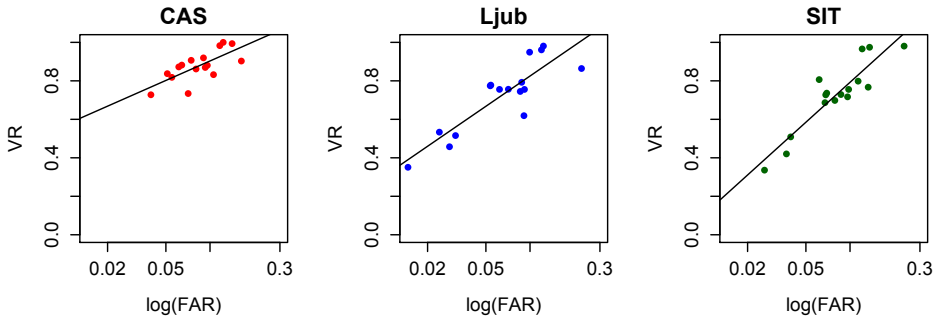


Figure 3. Scatterplots of VR vs log(FAR) of environment-pairs fitted with a linear regressor for each algorithm. Thresholds set to global FAR = 0.10.

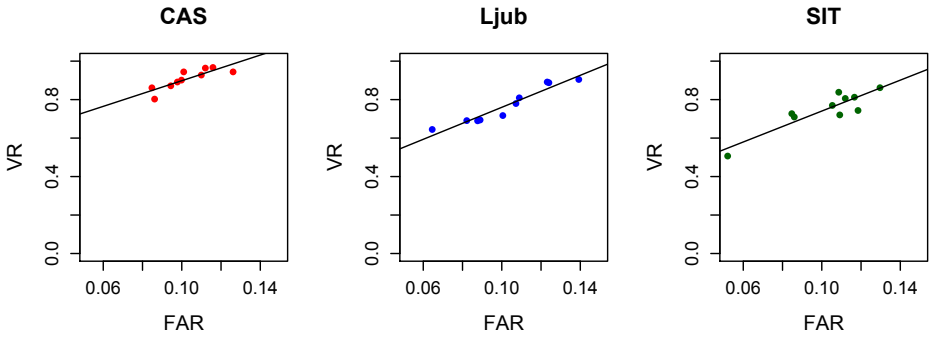


Figure 4. Scatterplots of VR vs FAR for Face Size, divided into 10 bins, fitted with a linear regressor for each algorithm. Thresholds set to global FAR = 0.10.

rial. Across all three algorithms for all three video factors, we saw a trade-off between VR and FAR for different levels of each factor. Further analysis suggested an interaction between environment and both face size and face confidence with face size having a larger interaction.

## 6. Results for Demographic Factors

It is known that gender and race effect the performance of algorithms [11], [5]. Figure 6 shows the effect of gender and race on the marginal FAR for Algorithm **Ljub**. The corresponding results for all three algorithms are reported in Figure 11 in the Supplemental Material. The results show that cross-gender and cross-race impostor-pairs have a lower FAR. This is consistent with O'Toole et al. [11].

## 7. Subject Identities

Subject identity as a factor has been studied fairly extensively, often under the heading "The Biometric Zoo" [4],

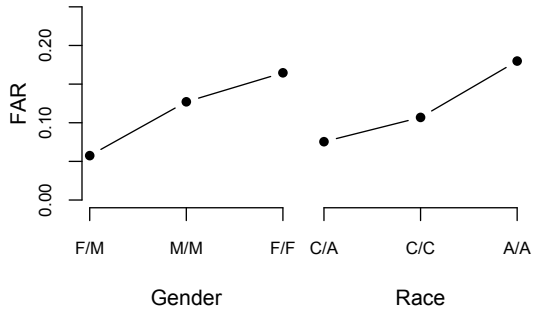


Figure 6. FAR for demographic factors for **Ljub**. The FAR for each factor-level is reported for a global FAR = 0.10. For gender, there are three factor levels: female-male (F/M), male-male (M/M), and female-female (F/F). For race, there are three factor-levels: Caucasian-Asian (C/A), Caucasian-Caucasian (C/C), and Asian-Asian (A/A).

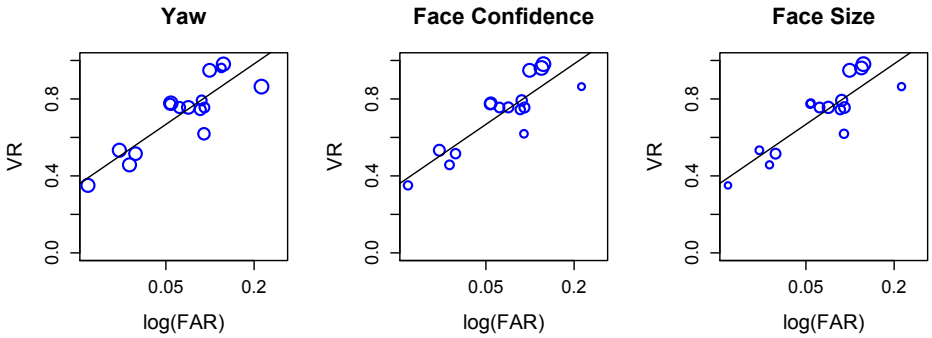


Figure 5. Interactions between Algorithm **Ljub** environment-pairs from Figure 3 and each of the three video-based factors: yaw, face confidence, and face size. Each panel looks at the interaction for the factor in its title. The size of each circle is proportional to the mean of the factor for each environment-pair.

[16]. However, defining factor levels based upon identity is problematic and so instead we move to the more interesting question of whether the marginal VR for a person correlates with the marginal FAR. In other words, do we see for people the same connection between VR and FAR as found for the other factors addressed above. To answer this question, for each algorithm, the 265 subjects are rank ordered by marginal FAR and marginal VR. Spearman's rank correlation coefficient for these tests are 0.15, 0.24 and 0.36 for algorithms **CAS**, **Ljub** and **SIT** respectively. In short, unlike the other factors studied, VR and FAR are not strongly correlated for people. This finding is consistent with previous zoo studies on unconstrained face recognition [14].

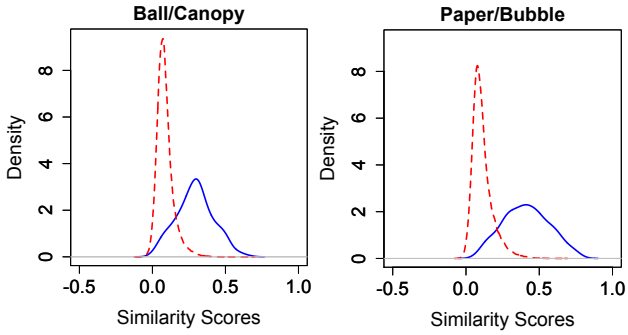


Figure 7. Density estimations of the match (solid blue) and impostor (dashed red) distributions of the environment pairs Ball/Canopy (Ba-Ca) and Paper/Bubble (Pa-Bu) for Algorithm **CAS**.

## 8. Impostor Distributions and Normalization

The variability in the marginal FAR clearly shows variability in the impostor distributions. The next question to ask: what type of variability? There are two distinct possibilities. First, the distribution related by an affine transformation; e.g., a shift in the location parameter and a change in the scale parameter. Alternatively, their differences are arising in the tails of the distributions with one distribution's tail substantially heavier than another.

Figure 7 shows density estimates of the match and impostor distributions for two environment-pairs. It appears that the tails of the impostor distributions are different. To get a better sense of differences in the tails of the impostor distributions, we will look at qq-plots.

A qq-plot is a graphical method to compare two distributions, particularly with respect to their skewness and tail behavior. In Figure 8, we compare the impostor distributions for two environment-pairs for Algorithm **CAS** . When the qq-plot shows a straight line, as it does for the left panel, it means that the two distributions are merely shifted, and possibly rescaled, versions of each other. The right panel shows a very interesting result where the impostor distributions for the Paper-Bubble (Pa-Bu) and the Ball-Canopy (Ba-Ca) environment-pairs have very different shapes: the upper (right) tail of the distribution for Ball-Canopy (Ba-Ca) impostor distribution is heavier and extends further than does the upper tail of the Paper-Bubble (Pa-Bu) impostor distribution. The two impostor distributions have fundamentally different shapes: Ball-Canopy (Ba-Ca) has many more large impostor scores and more extreme ones.

We found that the majority of qq-plots indicated an affine relationship (i.e., shift and scale) among the impostor distributions. There were a number of environment-pairs where the impostor distribution shapes were different; e.g., not affine. The generalizability across algorithms for affine cases was unexpected.

We now proceed with a quantitive experiment to test the conclusions of our exploratory data analysis on the shape of

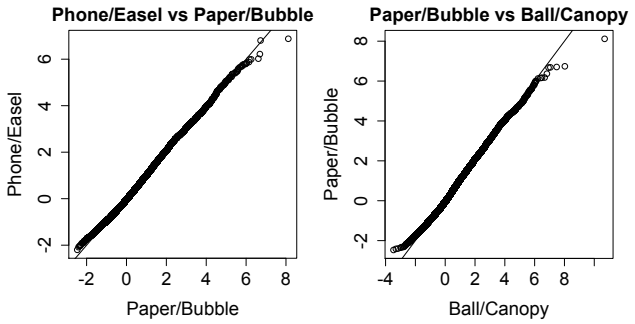


Figure 8. QQ-plots comparing environment pair impostor distributions of Phone/Easel (Ph-Ea) to Paper/Bubble (Pa-Bu) and of Paper/Bubble (Pa-Bu) to Ball/Canopy (Ba-Ca) from Algorithm **CAS**. In the left panel, the distributions are simply shifted and scaled versions of each other while in the right panel, the distributions are different.

the impostor distributions. If the impostor distributions for the 16 cross-week environment-pairs are all affine transformations of each other, then it should be possible to normalize the distributions to be the same. If the normalized distributions were the same, then the FAR over the 16 environment-pairs should be the same. We experimented with two normalization methods. Independently, for each environment-pair, we z-normed its impostor distribution to have mean 0 and standard deviation 1. From all 16 z-normed impostor distributions, we computed a threshold $\tau_g$ for a global FAR = 0.10. Using the new threshold, we calculated the FAR for the environment-pairs and then computed the standard deviation over the 16 FARs. If the standard deviation for the z-normed distribution was substantially smaller than the standard deviation for the unnormalized distributions, then we assume that the impostor distributions were affine transformations of each other. The results of this experiment are in Table 2. We repeated the experiment with a robust normalization that shifted the distributions by the median and scaled by the median absolute deviation (MAD).

Table 2. The standard deviation (SD) of the FAR for the 16 cross-week environment-pairs. The standard deviation is reported for three conditions. Unnormalized: the impostor distribution is not normalized; z-norm: the distribution is shifted by the mean and scaled by the SD; and robust norm: the distribution is shifted by the median and scaled by the MAD.

| Algorithm | Unnormalized | z-norm | Robust norm |
|---|---|---|---|
| **CAS** | 0.034 | 0.047 | 0.012 |
| **Ljub** | 0.051 | 0.003 | 0.005 |
| **SIT** | 0.050 | 0.004 | 0.005 |

For Algorithms **Ljub** and **SIT**, the z-norm was effective, which suggests that the impostor distribution are affine translations of each other. The z-norm and robust normalization were equally effective for Algorithms **Ljub** and **SIT**. For Algorithm **CAS**, robust normalization was substantially better than the z-norm, but did not reduce the standard deviations as much as for the other two algorithms. This suggests that there is variability in the tails of the impostor distribution.

## 9. Conclusions

We have shown that environment and video factors effect the FAR for three algorithm on the video portion of the PaSC face recognition challenge. Surprisingly, for environment and video-based factors there was a clear relationship between VR and FAR. For these factors, one level is not better than another; there is a trade-off between VR and FAR. An increase (resp. decrease) in the FAR results in an increase (resp. decrease) in the VR. Also, unexpectedly, impostor distributions in most cases undergo simple translation and scaling when shifting between factor levels. In only a few cases is the change more complex. Our results illuminate a path for better understanding the performance of face recognition algorithms in unconstrained scenarios. The results underscore a need to better control a tendency of current algorithms to increase impostor scores in favorable settings as defined by higher true-match scores. These results also establish a foundation for better modeling of distributional changes conditioned on measurable, knowable, attributes of target application environments, and consequently bring us closer to the goal of predicting performance in new settings.

## References

[1] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, and Y. M. Lui. Focus on quality, predicting FRVT 2006 performance. In *Proceeding of the Eighth International Conference on Automatic Face and Gesture Recognition*, 2008. 2, 4, 5

[2] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *IEEE Conference on Biometrics: Theory, Applications and Systems*, 2013. 2

[3] J. R. Beveridge, H. Zhang, B. A. Draper, P. J. Flynn, Z. Feng, P. Huber, J. Kittler, Z. Huang, S. Li, Y. Li, M. Kan, R. Wang, S. Shan, X. Chen, H. Li, G. Hua, V. Štruc, J. Križaj C. Ding, D. Tao, and P. J. Phillips. Report on the FG 2015 video person recognition evaluation. In *Proceedings Eleventh IEEE International Conference on Automatic Face and Gesture Recognition*, 2015. 3

[4] G. Doddington, W. Ligget, A. Martin, M. Przybocki, and D. Reynolds. Sheeps, goats, lambs, and wolves: A statistical

analysis of speaker performance in the NIST 1998 recognition evaluation. In *Proceedings ICSLP '98*, 1998. 6

[5] G. H. Givens, J. R. Beveridge, P. J. Phillips, B. A. Draper, Y. M. Lui, and D. S. Bolme. Introduction to face recognition and evaluation of algorithm performance. *Computational Statistics and Data Analysis*, 67:236–247, 2013. 2, 6

[6] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807 – 813, 2010. 1

[7] 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. 3

[8] Y. Lee, P. J. Phillips, J. J. Filliben, J. R. Beveridge, and H. Zhang. Generalizing face quality and factor measures to video. In *International Joint Conference on Biometrics (IJCB)*, 2014. 1, 2, 4, 5

[9] 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. 3

[10] 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. 3

[11] A. J. O'Toole, P. J. Phillips, X. An, and J. Dunlop. Demographic effects on estimates of automatic face recognition performance. *Image and Vision Computing*, 30:169–176, 2012. 2, 6

[12] 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, and the ugly face recognition challenge problem. In *Proceedings Ninth IEEE International Conference on Automatic Face and Gesture Recognition*, 2011. 2

[13] A. Sgroi, K. W. Bowyer, P. Flynn, and P. J. Phillips. SNoW: understanding the causes of strong, neutral, and weak face impostor pairs. In *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013. 2, 5

[14] M. N. Teli, J. R. Beveridge, P. J. Phillips, G. H. Givens, D. S. Bolme, and B. A. Draper. Biometric zoos: Theory and experimental evidence. In *International Joint Conference on Biometrics*, 2011. 7

[15] J. K. V. Štruc and S. Dobrišek. MODEST face recognition. In *International Workshop on Biometrics and Forensics (IWBF'15) (under review)*, 2015. 3

[16] N. Yager and T. Dunstone. The biometric menagerie. *IEEE Trans. Pattern Analysis Machine Intelligence*, 32(2):220–230, 2010. 7

# Supplemental Material

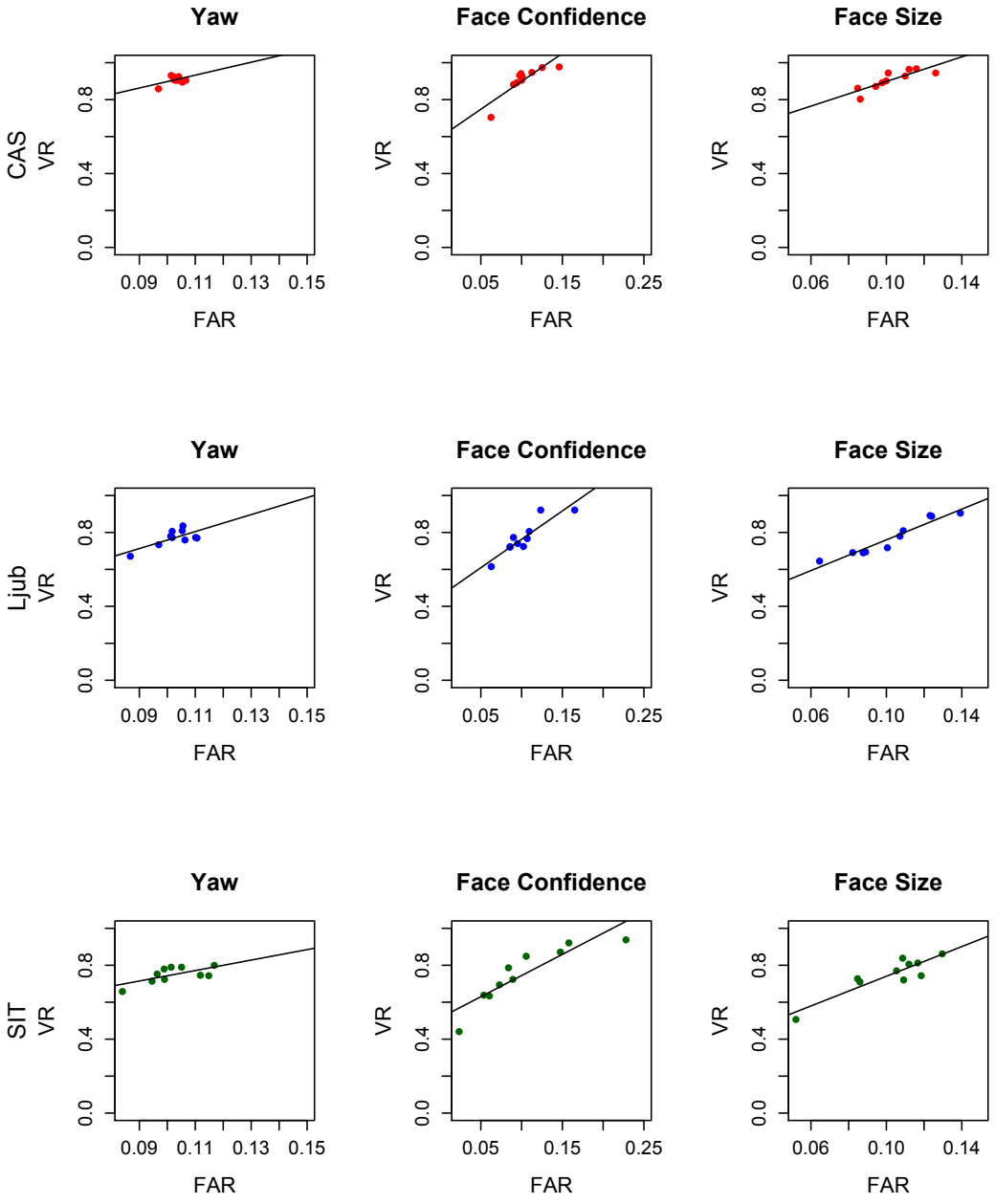


Figure 9. Scatterplots of VR vs FAR for video-based factors, fitted with a linear regressor for each algorithm. There are nine scatterplots, one for each algorithm and video-based factor. One column for each video factor and one row for each algorithm. All video-based factors are divided into 10 bins. Thresholds set to global FAR = 0.10.

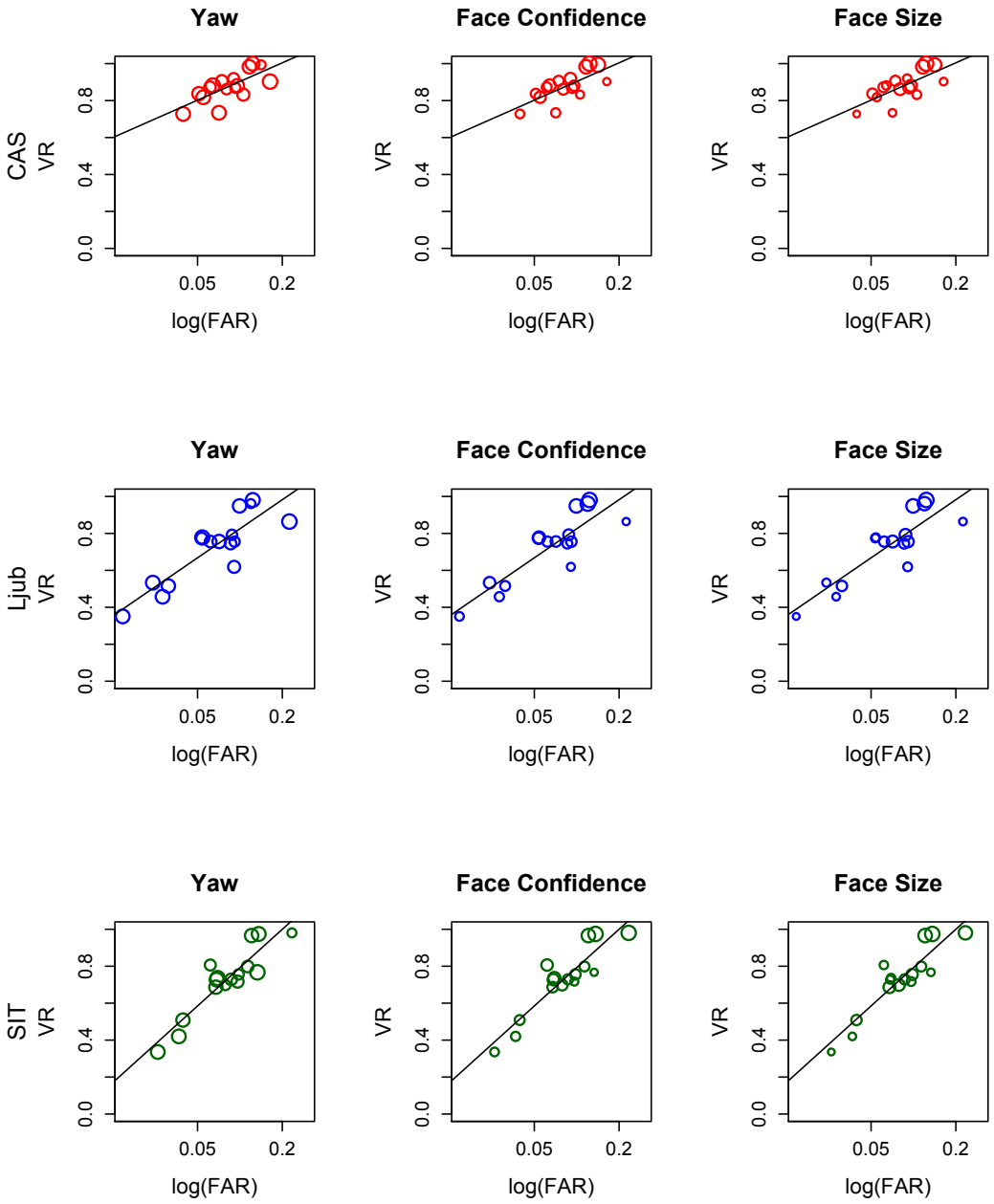


Figure 10. Interactions between environment-pairs and video-based factors. There are nine scatterplots, one for each algorithm and video-based factor. One column for each video factor and one row for each algorithm. Each panel looks at the interaction for an algorithm between environment-pairs and the factor in its title. The size of each circle is proportional to the mean of the video factor for each environment-pair.

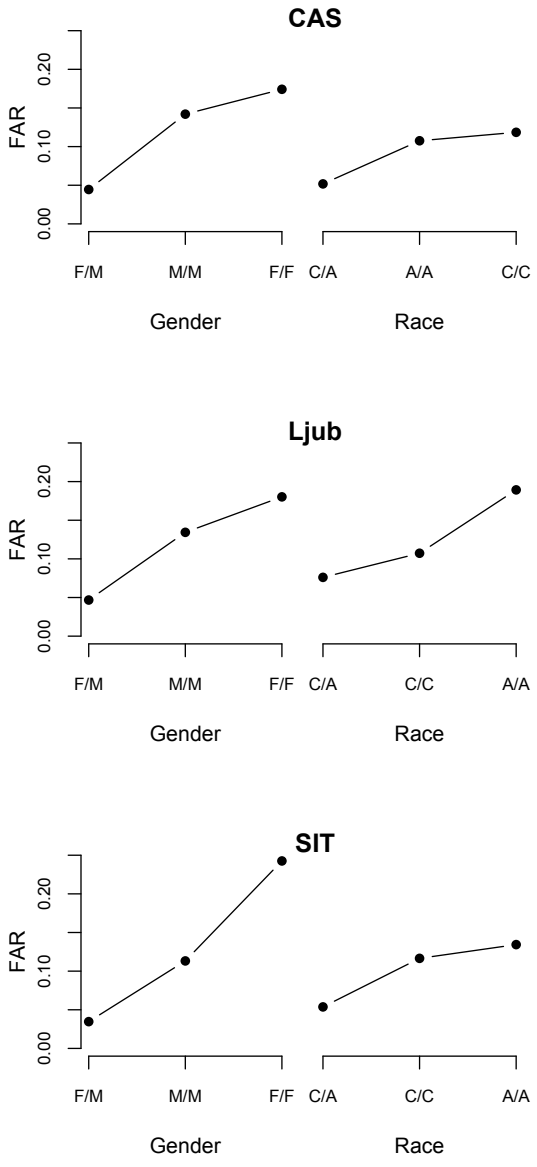


Figure 11. FAR for demographic factors for Algorithms **CAS**, **Ljub**, and **SIT**. The FAR for each factor-level is reported for a global FAR = 0.10. For gender, there are three factor levels: female-male (F/M), male-male (M/M), and female-female (F/F). For race, there are three factor-levels: Caucasian-Asian (C/A), Caucasian-Caucasian (C/C), and Asian-Asian (A/A).