# Predicting Face Recognition Performance in Unconstrained Environments

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

J. Ross Beveridge
Department of Computer Science, Colorado State University
Fort Collins, CO, USA

Geof Givens
Givens Statistical Solutions, LLC
Fort Collins, CO, USA

## Abstract

*While face recognition algorithms perform under many different unconstrained conditions, predicting this performance is not possible when a new location is introduced. Analyzing the impostor distribution of the videos of the Point-and-Shoot Challenge (PaSC) as well as its relationship to the genuine match distribution, we show that there is large variation in the false accept rate over the impostor distribution, demonstrate there is a correlation between changes in the verification and false accept rates over factor, and using this, present a method for predicting the performance of an algorithm using only unlabeled data for a new location.*

## 1. Introduction

Face recognition algorithms operate under a variety of unconstrained conditions, and performance varies substantially across locations, i.e., the physical location. Given videos in a new location, how well will an algorithm perform? Without explicitly testing the new location, a common method is to use the overall performance of the system on previously known locations. We present a way to better model the performance without needing to identify and label individuals in the videos.

Are there any locations that are "easy" (high verification rate and low false accept rate)? It is commonly believed that there exist locations that are easy as well as ones that are "hard." From our analysis, we show that such locations do not necessarily exist.

On the Point-and-Shoot Face Recognition Challenge (PaSC) [2], Lee et al. [10] found that verification rate (VR) varies across locations. The effect of factors on the genuine match distribution has been studied [1], [7]. However, there has not been as much research on the impostor distribution. O'Toole et al. [13] found that performance changes when the impostor distribution is restricted to people of the same gender or race. Several researchers have focused on the effects of pose, expression, and illumination [8], [6].

Extending the work of Lee et al. [10] , we investigate how the false accept rate (FAR) varies across locations in the PaSC dataset. The algorithms in this study are from the Face and Gesture 2015 Person Recognition Evaluation [3]. In our analysis, we include video-based factors which are automatically computed [10]. We also analyze the relationship between the genuine match distribution and impostor distribution. Using this analysis, we demonstrate it is possible to predict the performance of an algorithm in a new location based solely on unlabeled data acquired from the new location.

Novel contributions in this paper are:

- We show that when a threshold is set so that the global FAR is fixed, there is a large variation in the false accept rates over the locations.
- We show that with this fixed threshold, changes in verification and face accept are correlated across locations.
- This correlation allows us to predict the verification rate for new locations using a regression model.

## 2. PaSC Challenge and Data Set

To investigate the false accept rate across the impostor distribution, we needed a data set that documented many factors about the videos themselves, especially with the location of videos systematically varied. The Point-and-Shoot Face Recognition Challenge (PaSC) was designed to advance the development of face recognition algorithms on videos taken with digital point and shoot cameras, particularly for handheld cameras found in cell phones; full details of the protocol can be found in [2]. What follows is a brief summarization of the relevant details.

## 2.1. Data Set

In our analysis, we focus on the effect of location and sensor. This is possible because videos in the PaSC are taken from six locations with six sensors, five of those being handheld.

In the video portion of the PaSC, 2802 videos of 265 subjects were taken over 7 different weeks at the University of Notre Dame in the spring semester of 2011. 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.

Each subject is present in videos for at least four of the weeks, implying the differences in weeks' performances is not due to the subjects. Video length ranges roughly between 50 and 400 frames with most videos containing between 200 and 250 frames, and the resolutions ranged between 640×480 to 1280×720.

There were six different locations with six different sensors. Five of the sensors were handheld, and these varied by week. Additionally, data was collected by a tripod-mounted sensor, and this sensor filmed the same actions at the same location and time as the handheld sensor of the week.

Figure 1. Sampled portions of video frames from PaSC videos indicating some of the situations that make recognition challenging. Courtesy of Beveridge et al. [4].

Figure 1 shows a sample of frames from PaSC videos from different locations. Characterizing the videos are four primary factors: location, action being performed, video camera (sensor), and person in the video (subject).

## 2.2. Location Factor

One aspect the design of this data set allows us to analyze is how an algorithm performs when restricted to pairs of videos from certain locations. During each week, the videos were collected with a new combination of location and action taking place, for example picking up a newspaper in an office. No combination of location and action was repeated on subsequent weeks. Table 1 shows a summary of the location, handheld camera, and action combinations.[1]

Table 1. Location, camera, and action combinations. The abbreviations for the location 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 |

Each location and action combination was captured on a specific week by two different cameras, one being handheld. Consequently, each video depicts a single subject at a certain location doing a specific action captured by one particular sensor, e.g. for a specific subject, there is exactly one video depicting the subject on the lawn blowing bubbles captured by a Sanyo Xacti. There is also a video of the subject blowing bubbles on the lawn captured by the tripod-mounted sensor, a Panasonic HD700. From Table 1, it is clear that the handheld sensors are confounded with the locations and actions.

In the findings below, the influence that location, camera, and action combinations (called the location factor for simplicity) 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 was 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.

As videos are compared in pairs, the location factor is defined by location-pairs, i.e. the locations of the videos for a given pair. In total, there are 22 location-pairs. For 6 pairings the videos are from the same location and collected in the same week; these only include impostor pairs, i.e. pairs

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

of videos of different people. However, we focus mainly on cross-week comparisons, i.e. video-pairs in which the weeks of capture are different. There are 16 cross-week location-pairs. In 15 of the cross-week location-pairs, the videos were collected at different locations from different weeks, but for one pair, the videos were collected at the same location (canopy) on different weeks.

### 2.3. Video-Based Factors

Location, action, and sensor are not the only factors effecting performance. Another class of factors effecting performance comes directly from the videos themselves; that is, these factors, called video-based factors, are dependent on the video from which they are estimated. As we show later in Section 6, video-based factors can encode properties of a location-pair. For our work, we measure this encoding by looking at aggregate statistics of video-based factors from all video-pairs of the location-pair.

We consider three video-based factors: face size, face confidence, and yaw. Estimated by the Pittsburgh Pattern Recognition (PittPatt) face recognition SDK 5.2.2, face size is the number of pixels between the eyes, face confidence is PittPatt's self-assessment of how certain the algorithm was in detecting the true face, and yaw is the measurement of how far the face was turned to the left or right.

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 PittPatt SDK 5.2.2 software estimated these factors for the frames of the videos, and the generalizations to videos and video-pairs follow the methods of Lee et al. [10].

## 3. Algorithms

Our analysis is performed on the four top performers in the Face and Gesture 2015 Person Recognition Evaluation [3]. The algorithms were developed independently by four different research groups from four different countries on four different continents. Each algorithm is very different in how it computes a similarity score (the degree of similarity between two faces in two videos). This independence provides evidence that our conclusion will generalize to algorithms not included in this study.

The Chinese Academy of Science (CAS) algorithm uses two convolutional neural networks, one for larger and one for smaller faces [9].

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 [11], [12].

The University of Ljubljana (Ljub) algorithm combines four feature types with a probabilistic principal component analysis [15].

The Univeristy of Technology, Sydney, (UTS) algorithm uses three-dimensional face pose normalization and face descriptors [5].

## 4. Measuring Performance

Our results are reported on participants in the Face and Gesture 2015 Person Recognition Evaluation [3], and in this competition, the participants followed the PaSC protocol. In the protocol for the PaSC, algorithms are given two videos and then return a number measuring the degree of similarity between the subjects in the pair of videos. Hence, in calculating and predicting performance, we compare videos in pairs.

In measuring performance, we are observing how often an algorithm correctly declares the same person to be in two videos. We are also interested in how often the algorithm incorrectly believes two different people from videos are the same person. However, we are not interested in the overall performance of the algorithm. Instead, we are more interested in how the performance changes over levels of a factor. Later in this paper, for a set of videos of a factor-level, we are predicting how well an algorithm will correctly match videos of the same person (marginal VR). In our prediction, we use how often the algorithm incorrectly declared different people to be the same (marginal FAR). We then compare our predicted performance to the actual observed performance.

The focus of analysis in this paper is on performance when comparing videos for a factor-level. Presented with two faces from videos $x$ and $y$, an algorithm $A$ returns a similarity score, $s_A(x, y)$, for video-pair $(x, y)$. The similarity score denotes how similar the faces are estimated to be; a higher similarity score indicates a higher likelihood of the two faces belonging to the same subject.

To make a decision, a threshold $\tau_g$ is set so that every video-pair score at least as large $\tau_g$ is declared a match and every score below the threshold is considered a non-match. We divide the set of videos into two sets: the set of video-pairs that are genuine matches and the set of video-pairs that are impostors. With the threshold $\tau_g$, we calculate the verification rate $\mathrm{VR}(\tau_g)$ as the ratio of genuine matched video-pairs correctly identified as a match and the false accept rate $\mathrm{FAR}(\tau_g)$ as the ratio of impostor video-pairs incorrectly identified as a match.

Generally, the threshold $\tau_g$ is set to specify the FAR at a certain instance. In our paper, we select $\tau_g$ for each algorithm so that globally $\mathrm{FAR}(\tau_g) = 0.10$. For PaSC, the standard for reporting VR is $\mathrm{FAR} = 0.01$. However, we shifted the threshold to have enough false matches for analysis.

Nonetheless, the analysis in this paper is not focused on the overall performance over the set of all video-pairs.

Rather, for this paper, as previously mentioned, the analysis is centered on performance when comparing video-pairs of factor levels such as locations. For the marginal verification and false accept rates for factor $F_i$, a threshold $\tau_g$ is set so that globally $\mathrm{FAR}(\tau_g) = 0.10$, and with this threshold, the verification rate $\mathrm{VR}(F_i, \tau_g)$ and false accept rate $\mathrm{FAR}(F_i, \tau_g)$ are then calculated only on the video-pairs in $F_i$.

## 5. Imposter-Pair Analysis For Location-Pairs

In this section, we picked the threshold $\tau_g$ so that the global FAR = 0.10 and then investigated the impostor distribution over the different location-pairs, calculating the marginal false accept rates over the location-pairs using the threshold $\tau_g$. We showed that there is large variation in the false accept rate. Then we showed that, keeping the threshold $\tau_g$ constant, the changes in the verification and false accept rates over the location-pairs are correlated.

### 5.1. Range of Marginal FARs over Location-Pairs

It is well known that location significantly effects algorithm performance. The design of the PaSC data set enabled us to characterize the impact of location on performance. Previous studies have investigated the effect of location on verification rates [1], [10]. We proceed by examining the effect of location on the FAR and then look at the relationship between FAR and VR.

Since comparisons are between two videos, we look at performance for location-pairs. For the four algorithms in our study, we computed the FAR for the 22 location-pairs as described in Section 4. Figure 2 demonstrates how location factors effect FAR (upper graph) and VR (lower) for the four algorithms on handheld video-pairs when the global FAR is set to 0.10. Along the horizontal axes are the pairs of locations described in Section 2.2. All 22 pairs are present in the upper graph, but only the 16 cross-week pairs are present in the lower graph because the same-week comparisons only contain impostor pairs. The vertical axes show the marginal FAR and VR values, respectively, using a $\tau_g$ that corresponds to a global FAR of 0.10. The location pairs are ordered by the mean rate over all the algorithms for both graphs. In the top graph, all location pairs to the left of the 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 location exerts a dramatic influence over the impostor distribution and hence the marginal FAR. For handheld video-pairs, Algorithm Ljub has the greatest range in FAR from 0.01 to 0.42, and CAS has the smallest range from 0.05 to 0.27; for tripod video-pairs, Ljub still has the greatest range in FAR from 0.02 to 0.39, and CAS has the smallest range from 0.03 to 0.22. Table 2 shows the ranges for the cross-week location-pairs over both sets of video-pairs. For the handheld video-pairs, the FAR for the four algorithms CAS, UTS, Ljub, and SIT varies by a factor of 3.6, 7.33, 21, and 11.5, respectively. For the tripod video-pairs, the FAR for the algorithms CAS, UTS, Ljub, and SIT varies by a factor of 4.33, 7.67, 9, and 7, respectively. Prior work has already suggested the importance of location [1], [10]; this is the first clear evidence of how significantly it effects the impostor distribution.

Table 2. The cross-week ranges of location-pair marginal $\mathrm{FAR}(L_i, \tau_g)$ location-pairs over both sets of video-pairs with a threshold $\tau_g$ set so that global FAR = 0.10.

| Algorithm | Handheld | Tripod |
|---|---|---|
| CAS | $0.05 - 0.18$ | $0.03 - 0.13$ |
| UTS | $0.03 - 0.22$ | $0.03 - 0.23$ |
| Ljub | $0.01 - 0.21$ | $0.02 - 0.18$ |
| SIT | $0.02 - 0.23$ | $0.03 - 0.21$ |

A related finding is the importance of the cross-week versus same-week distinction. For both sets of video-pairs, the mean cross-week marginal FAR averaged over the algorithms was 0.09 compared to 0.21 for same-week pairs. A recent related result on still face image by Sgori et al. [14] 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.

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

We will now look at the relationship between the location-pair FARs and VRs for the cross-week pairs. Scatterplots in Figure 3 relate marginal VR to marginal FAR, described in Section 4, for the 16 cross-week location-pairs over the different sensor-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 location-pairs over different sensor-pairs, and the line is a linear regressor. For all four algorithms, the regression line suggests a linear relationship between $\log(\mathrm{FAR})$ and $\mathrm{VR}$. In other words, a location-pair that has a higher marginal VR will likely have a higher marginal FAR. Unfortunately, this linear relationship suggests that finding a location-pair that is easier than others is unlikely. We say a location-pair is easier if it has both a higher VR and a lower FAR than other pairs.

## 6. Imposter-Pair Analysis For Video-Based Factors

In this section, we investigated the impostor distribution over the different video-based factors and showed that there is large variation in the false accept rate. Then we

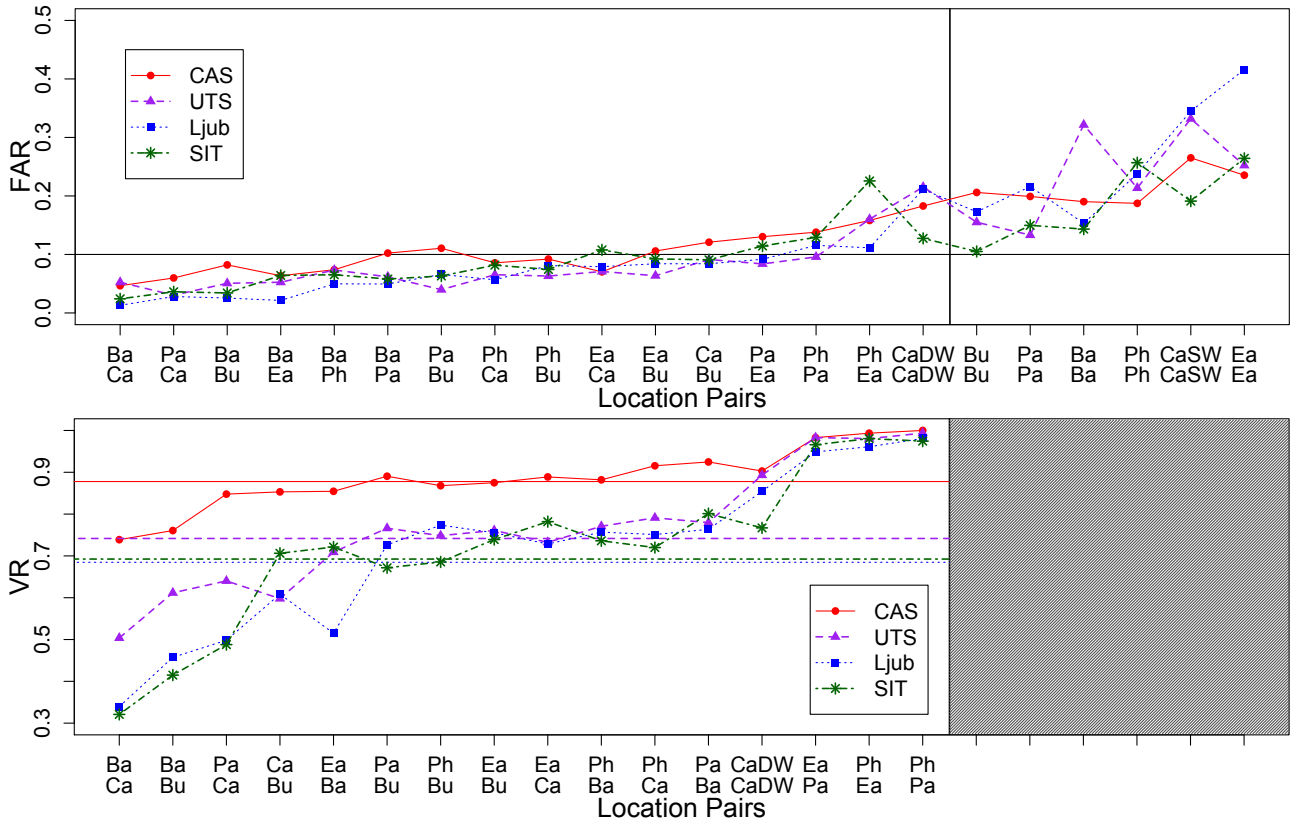


Figure 2. With a threshold set $\tau_g$ so that the global FAR is fixed, these graphs show the marginal $\mathrm{FAR}(L_i, \tau_g)$ and $\mathrm{VR}(L_i, \tau_g)$ of each location-pair on handheld video-pairs for each algorithm–ordered by the mean rate over all the algorithms. The top graph is on $\mathrm{FAR}(L_i, \tau_g)$. The horizontal line corresponds to the global FAR = 0.10, and the vertical line between pairs CaDW-CaDW and Bu-Bu separates the pairs into cross-week (left) and same week. The bottom graph is on $\mathrm{VR}(L_i, \tau_g)$. The horizontal lines correspond to the global $\mathrm{VR}(\tau_g)$ for each algorithm when the global FAR = 0.10. There are no same-week pairs for matches.

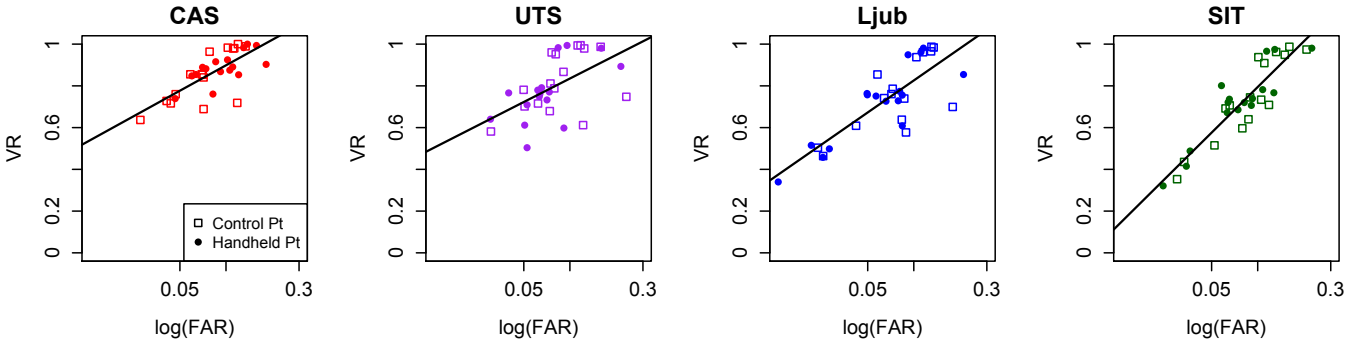


Figure 3. Scatterplots of $\mathrm{VR}(L_i, \tau_g)$ vs $\log(\mathrm{FAR}(L_i, \tau_g))$ of location-pairs over different sensor-pairs with a threshold $\tau_g$ set to that global FAR = 0.10.

showed that changes in the verification and false accept rates over the location-pairs are correlated and interact with the location-pairs.

The impact of image- and video-based factors on verification rates has 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 location-pairs and the video-based factors.

Figure 4 shows the trade-off between FAR and VR for face size. The procedure described at the end of Section 2.3 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 location 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.

Figure 5 highlights the interactions between location and video factors for Algorithm Ljub. Like the scatterplots in Figure 3, each point corresponds to a location-pair and sensor-pair. Unlike in Figure 3, in Figure 5 circle size varies and is proportional the mean video factor for a location-pair. For the yaw-factor, all the circles are about the same size, which means that yaw does not interact with the location-pair. In contrast, a clear interaction effect between location and face size is evident: location-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 location and face confidence.

This analysis was repeated for Algorithms SIT, UTS, and CAS, and the conclusions were the same. Across all four 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 location and both face size and face confidence with face size having a larger interaction.

## 7. Predicting Performance

### 7.1. Models

With a new, previously unseen, location being compared to a known location, how well can performance (marginal VR) be predicted? We know that there is a wide range of potential marginal VR. Figure 3 illustrates this, showing scatterplots of $\mathrm{VR}$ vs $\log(\mathrm{FAR})$ of location-pairs over different sensor-pairs. Recall that additionally, a linear regressor is fit to the points for each algorithm. Observe the ranges of the marginal VR for the location-pairs of the four algorithms. For the algorithm SIT, the range is from 0.32 to 0.99 when the threshold $\tau_g$ is picked to set the global FAR to 0.10, described in Section 4.

What if, instead of one new location, two locations are new and compared against each other? How well can we accurately predict performance of this entirely new pair? Is it even possible to predict the performance with the same technique used when only one location is new? Which factors should be included in a model?

We started with a very simple model. As explained below, Linear Model 1 uses only the $\mathrm{FAR}$ of a location-pair to predict what the observed $\mathrm{VR}$ will be. Simply knowing how many false positives are in the set of video-pairs for a location-pair can indicate how well the algorithm will perform for those video-pairs. Additionally knowing some more information on the video-pairs, i.e. the video-based factors from Section 2.3, a better prediction can be made using Linear Model 2.

In Figure 3, a simple linear regressor is fit solely to the marginal verification and false accept rates of the location-pairs. The linear regressor is given by

$$\mathrm{VR} = \alpha + \beta \log(\mathrm{FAR}). \tag{1}$$

This is Linear Model 1.

Video-based factors are not incorporated into Linear Model 1. However, as we noted earlier, there is interaction between location and two video-based factors. There is interaction between location and face size, there is less interaction between location and face confidence, but there is no interaction seen between location and yaw.

To find a second model that utilizes video-based factors, we removed each location and partitioned the subjects into training and testing sets. On the remaining video-pairs that had both subjects in the training set, we fit models on the marginal $\mathrm{VR}$ using marginal $\mathrm{FAR}$ as well video-based factors from Section 2.3 and any relevant two-way interaction terms for each location-pair; we only kept terms that were significant ($p < 0.05$).

Many models resulted, and they performed robustly the same across the algorithms indicating that specifically which terms are in the model is not highly significant. With a set of second models being robustly the same in terms of prediction performance, we chose for Linear Model 2 to be given by

$$\mathrm{VR} = \alpha + \beta_1 \log(\mathrm{FAR}) + \beta_2 \,\mathrm{Yaw} + \beta_3 \,\mathrm{FC} + \beta_4 \,\mathrm{Yaw} * \log(\mathrm{FAR}) \tag{2}$$

where $\mathrm{Yaw}$ is the mean yaw and $\mathrm{FC}$ stands for the mean face confidence for the video-pairs of the location-pair. We use these models in the method described below in Section 7.2 for predicting performance.

### 7.2. Prediction Procedure

In order to predict how well a set of videos of a location-pair might perform, we do the following. There are sixteen cross-week location-pairs over different sensor-pairs. For each location-pair $L_i$, one of the locations is randomly dropped. There will be no location-pair (no video) containing the dropped location; this location will be new. On the video-pairs of the remaining cross-week location-pairs, the subjects are partitioned into two sets: training and testing. Only video-pairs with both subjects in the training set are used.

With the video-pairs of the training set subjects, the global threshold $\tau_g$ is set so that the global FAR is 0.10. The global VR is calculated over all video-pairs in the training set using $\tau_g$; this is denoted as $\mathrm{VR}_g$. For the extant location-pairs, none of which use the new location, the marginal val-

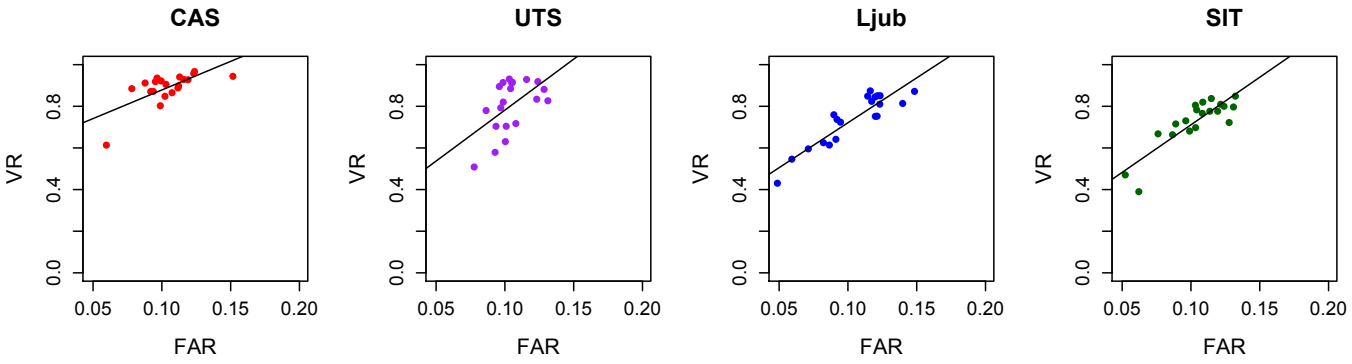


Figure 4. Scatterplots of $\mathrm{VR}(F_i, \tau_g)$ vs $\mathrm{FAR}(F_i, \tau_g)$ for Face Size over different sensor-pairs, divided into 10 bins, fitted with a linear regressor for each algorithm. Thresholds $\tau_g$ set to global FAR = 0.10.

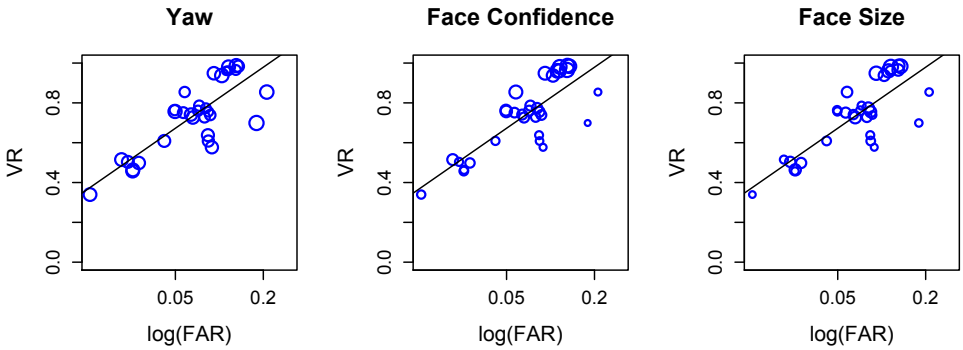


Figure 5. With threshold $\tau_g$ set for a global FAR = 0.10, interactions between Algorithm Ljub location-pairs from Figure 3 and each of the three video-based factors: yaw, face confidence, and face size. Each panel looks at the interaction for the factor in its title. The size of each circle is proportional to the mean of the factor for each location-pair.

ues are calculated over the different sensor-pairs, and these are used to fit the regression models from Section 7.1.

Using $\tau_g$ and the method described in Section 4, the observed marginal VRs of the location-pair $L_i$ are calculated over sensor-pairs; we denote this by $\mathrm{vr}_i$. Furthermore, the marginal FARs, $\mathrm{far}_i$, are also calculated. With the marginal values, a regression line can predict the observed verification rate. This predicted VR is $\widehat{\mathrm{vr}}_i = f(\mathrm{far}_i)$ where the function $f$ is Linear Model 1 (eq. 1) or Linear Model 2 (eq. 2).

The root mean square error (RMSE) is used to determine the standard deviation between the predicted VR and the observed VR ($\mathrm{vr}_i$). When using the global rate, $\mathrm{VR}_g$, to predict the observed VR, the RMSE is denoted by $\mathcal{G}$. When using the VR predicted by a regression line, $\widehat{\mathrm{vr}}_i$, the RMSE is denoted by $\mathcal{E}$. Equations 3 and 4 formally express the definitions, respectively.

$$\mathcal{G} = \sqrt{\frac{\sum_{i=1}^{n}(\mathrm{VR}_g - \mathrm{vr}_i)^2}{n}} \tag{3}$$

$$\mathcal{E} = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{\mathrm{vr}}_i - \mathrm{vr}_i)^2}{n}} \tag{4}$$

## 8. Results of Prediction

Are these models better than using the global VR? In order to test the models from Section 7.1, we implemented the procedure from Section 7.2 100 times with one location being new for each location-pair. Then, in order to test if the method was valid for two new, unseen locations, we ran the procedure another 100 times, but this time, both locations of a location-pair were new.

After 100 iterations of the Section 7.2 process, Figure 6(a) displays the mean RMSEs, equations 3 and 4, of predicting the observed VR with the previous global VR, with the VR produced from Linear Model 1, and with the VR produced from Linear Model 2 over all location-pairs and sensor-pairs. The bars extend one standard deviation. For Algorithms Ljub and SIT, the mean RMSEs from forecasting using Linear Model 1 are much lower than using the global VR, which are over 0.21. For the algorithms CAS

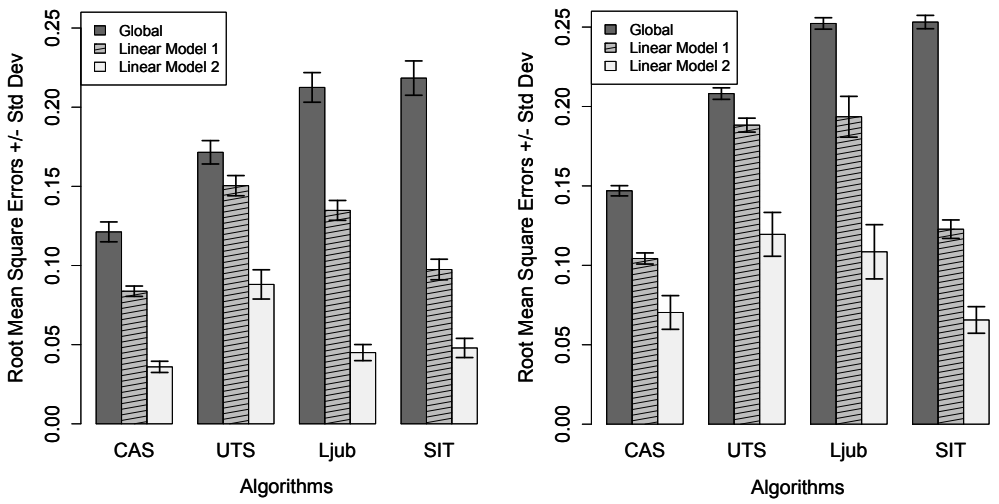


Figure 6. Bar plots of the mean RMSEs with standard deviation bars. In (a), one location is new. In (b), two locations are new.

and UTS, using Linear Model 1 is still better than using the global VR, which have mean RMSEs over 0.12, but the gap is not as large as it is for the other two algorithms.

The second linear model predicts the observed VR even better than the first linear model. The RMSEs from Linear Model 2 are much smaller than those from Linear Model 1 and definitely from those using the global VR. In fact, the means from Linear Model 2 are below 0.05 across three of the algorithms: CAS, Ljub, and SIT. The mean RMSE of Algorithm UTS is under 0.09, which is much smaller than it was from using the global VR or Linear Model 1 VR.

After 100 iterations, Figure 6(b) displays the mean RMSEs of predicting the observed VR with the global VR, with the VR produced from Linear Model 1, and with the VR produced from Linear Model 2 as in Figure 6(a), but in Figure 6(b), instead of one location being new, now both locations are new. Again, in general forecasting with Linear Model 1 is better than simply using the global VR. Using the global VR, Algorithm CAS has a mean RMSE of 0.15, and UTS has a mean RMSE of over 0.20. Algorithms Ljub and SIT have mean RMSEs over 0.25. For the algorithm SIT, the mean RMSE of Linear Model 1 less than half the mean RMSE of the global VR prediction. For Algorithms CAS, UTS, and Ljub, Linear Model 1 is still better than the previous global VR, but the differences are not as large as it is for SIT.

The second linear model still does even better than the first. There is a little more variability than before, but that is not surprising as now both locations are new. The mean RMSEs are under 0.12 for Algorithms UTS and Ljub, and the mean RMSEs are below 0.08 for Algorithms CAS and SIT.

## 9. Conclusion

We have shown that it is possible to predict the performance of an algorithm on unseen videos at a new location. We demonstrated that using the previously-known global VR is not a very good estimate; there is a lot of variability in marginal VR across location-pairs. We presented two models for predicting the marginal VR of a new location. The first model uses only the marginal FAR, and the second uses the marginal FAR as well as two video-based factors: yaw and face confidence. Both methods are better than simply using the previous global VR, but the second model came the closest to predicting the observed VR. Given two new locations, the second model is much better than using the global VR. The algorithms on which we tested were from four different groups on four different continents, implying that our results will generalize well.

To develop these models, we looked at the effect of location-camera-action (simply called location) and video factors on the FAR. Surprisingly, for location 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. 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 genuine match scores. These results also establish a foundation for better modeling of distributional changes conditioned on measurable, knowable, attributes of target application locations, and thus bring us closer to the goal of predicting performance on unseen videos at new locations.

## 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] 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.

[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.

[4] J. R. Beveridge, H. Zhang, P. Flynn, Y. Lee, V. E. Liong, J. Lu, M. Angeloni, T. Pereira, H. Li, G. Hua, V. Struc, J. Krizaj, and P. J. Phillips. The IJCB 2014 PaSC video face and person recognition competition. In *Proceedings of the International Joint Conference on BIometrics*, 2014.

[5] C. Ding, C. Xu, and D. Tao. Multi-task pose-invariant face recognition. *IEEE Transactions on Image Processing*, 24(3):980–993, March 2015.

[6] A. Dutta, R. N. J. Veldhuis, and L. J. Spreeuwers. Predicting face recognition performance using image quality. *CoRR*, abs/1510.07119, 2015.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] V. Štruc, J. Križaj, and S. Dobrišek. Modest face recognition. In *3rd International Workshop on Biometrics and Forensics (IWBF 2015)*, pages 1–6, March 2015.