SNoW: Understanding the Causes of Strong, Neutral, and Weak Face Impostor Pairs

Amanda Sgroi, Kevin W. Bowyer, Patrick Flynn
University of Notre Dame
Notre Dame, IN
asgroi kwb flynn @nd.edu

P. Jonathon Phillips
NIST
Gaithersburg, MD
jonathon.phillips@nist.gov

Abstract

The Strong, Neutral, or Weak Face Impostor Pairs problem was generated to explore the causes and impact of impostor face pairs that span varying strengths of scores. We develop three partitions within the impostor distribution for a given algorithm. The Strong partition contains image pairs that are easy to categorize as impostors. The Neutral partition contains image pairs that are less easily categorized as impostors. The Weak partition contains image pairs that are likely to cause false positives. Three algorithms, and the fusion of their scores, were used to analyze the performance of these three partitions using the same set of authentic scores employed in the Face Recognition Vendor Test (FRVT) 2006 Challenge Dataset. The results of these experiments provide evidence that varying degrees of impostor scores impact the overall performance and thus the underlying causes of weak impostor pairs are worthy of further exploration.

1. Introduction

In the authentic score distribution, scores in the tail that overlaps the impostor distribution have the potential to cause false rejects. The image pairs that cause this scenario have been studied at great length in order to improve recognition rates. However, the scores that fall in the corresponding tail of the impostor distribution, which can potentially cause false accepts, have mostly been ignored. Analysis of these scores, and of the impostor distribution as whole, has the potential to help us understand how to characterize an authentic image pair that has the most impostor-like score, and similarly, impostor image pairs that have the most authentic-like score.

Some studies have aimed to identify certain aspects of the relationship between a subject's authentic and impostor distributions, such as in biometric zoos [10]. In Doddington's zoo, both lambs and wolves are identified as subjects who cause false positives for gallery and probe images, respectively [3]. In Yager and Dunstone's zoo, chameleons are identified as subjects that match well to themselves and others, whereas worms are those that match poorly against themselves yet relatively well to others [11]. Both of the classes have the ability to cause a false negatives. Although biometric zoos identify subjects that can cause errors in a recognition system, no explanation is given for why false accepts occur for a particular subject and often contrary evidence is found within the two zoo structures [10].

A few studies have focused solely on the impostor distribution. O'Toole et al. [6] explore how the demographic composition of a dataset's impostor pairs affects algorithm accuracy. In their work, only impostor scores that occurred between subjects of the same race or gender were considered in the impostor distribution. Given this constraint, they showed changes in recognition performance for partitions of the impostor distribution of variable difficulty with respect to the false reject rate. Evidence also suggested that the demographic composition of the impostor distribution affects the level of performance estimated for an algorithm and the choice of threshold for authentic or impostor decisions.

Several studies however have attempted to model the impostor distribution in order to dynamically determine the threshold between authentic and impostor scores. In [9] the authors use extreme value theory on the distribution generated by one probe image. Since this distribution contains at most one authentic comparison and many impostor comparisons it is easy to perform outlier detection to find the authentic score, and then adjust the recognition system threshold. A second study explores the biometrics zoo to develop a quality measure determined by how similar a probe image is to a population of impostors [4]. The authors develop a Uniqueness Based Nonmatch Estimate (UNE) which offers an improvement in the overall threshold used for determining whether a score belonged to the authentic or impostor distribution.

In the Good, the Bad, and the Ugly (GBU) Face Recog-

nition Challenge problem the authentic score distribution of the Face Recognition Vendor Test (FRVT) 2006 Dataset was partitioned based on the difficulty of recognition [7]. Since the genesis of the GBU problem, several studies have considered the design of a quality metric that may describe what makes an image easy or hard to recognize [2][1]. To date, the primary insights from these works seem to be that quality is inherent to an image pair rather than a single image, and that varying location of image acquisition plays a key role in the ease or difficulty of a match comparison.

This paper aims to establish a framework roughly analogous to that of the GBU, but focused on analyzing the impostor distribution. We create three partitions of impostor image pairs, based on how strongly their score identifies them as an impostor pair. Analyzing image pairs across these three categories should improve our understanding of the factors that lead to false matches.

The remainder of the paper is organized as follows. Section 2 discusses the partitioning method and constraints, contrasting them with those used in the GBU problem. Section 3 presents the performance of each partition for four different face recognition algorithms, using the same data set. Section 4 discusses the results and conclusions, as well as outlines directions for future work regarding the understanding of the impostor distribution.

2. The Strong, Neutral, or Weak Partitions

In the GBU Challenge Problem, three partitions were defined to study an authentic score's degree of difficulty within the authentic distribution [7]. The Good partition contained image pairs that were easy to recognize; the Bad partition contained image pairs that were moderately difficult to recognize; and the Ugly partition contained image pairs that were very difficult to recognize. In order to generate the three partitions, three constraints were employed - the distinct images constraint, the balanced subject count constraint, and the different days constraint. Through the use of the top three performing algorithms from the FRVT 2006 evaluation, the authors were able to describe the performance of these partitions. At a false accept rate of 0.001, verification rates of 0.98, 0.80, and 0.15 were found for the Good, Bad, and Ugly partitions respectively. The goal of this work was to encourage the development of face recognition algorithms that are robust to a broad range of conditions in frontal face images.

The Strong, Neutral, or Weak (SNoW) problem is based on three partitions of impostor image pairs. The Strong partition consists of pairs of face images of different people which are easy to determine as an impostor pair. The Neutral partition contains pairs of face images of different people which are likely an impostor pair. The Weak partition contains pairs of face image of different people which may be considered to be an authentic pair; these are often the image pairs which cause false matches in a recognition system. In this work, the partitions were constructed from the Notre Dame multi-biometric data set used in the FRVT 2006 evaluation [8]. This dataset consists of 9,308 total images of 569 subjects [1]. All the images are frontal still faces collected either outdoors with uncontrolled illumination or indoors with uncontrolled ambient illumination. The images were acquired with a 6 Mega-pixel Nikon D70 camera, and have a size of 3008 pixels by 2000 pixels, with an average of 190 pixels between the eyes [2]. Further, all images were taken between August 2004 and May 2005.

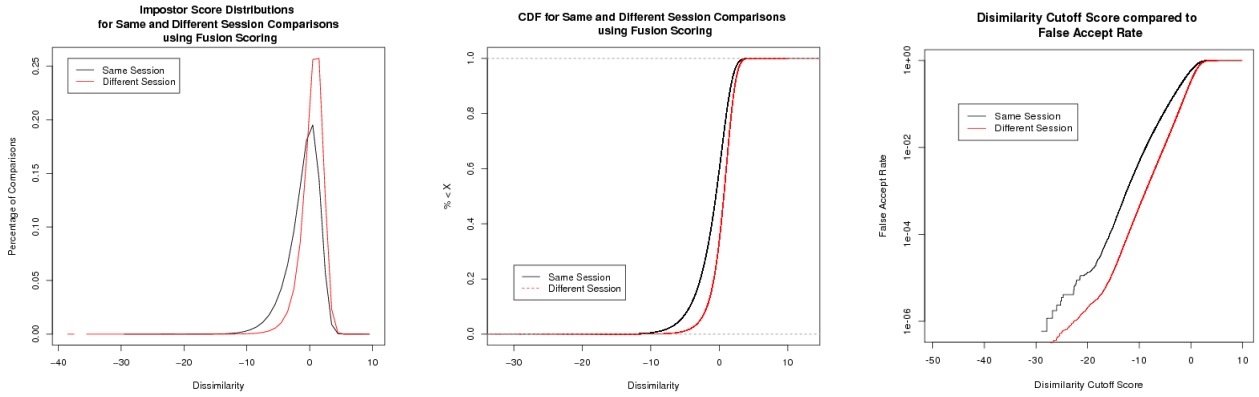
From the FRVT 2006 data, given a "symmetric" algorithm (producing the same score for a pair of images regardless of which one is considered the enrolled image), we can obtain 86,638,864 total scores. Of these scores, 225,286, or 0.26% are authentic scores, and 86,413,578, or 99.74%, are impostor scores. The impostor distribution as whole can be considered as the union of a set of impostor distributions for each subject. That is, for each subject, which is represented by some set of images, there is a set of scores produced from comparing these images to all images of all other subjects. This union of sets can be referred to as the **global impostor distribution** S. Then, the set of impostor pairs between a subject i, and all other subjects can be referred to as a **subject specific impostor distribution** S_i. Further, each subject pair defines an impostor distribution. These sets contain only impostor comparisons between a subject i and a subject j, referred to here as **subject pair impostor distributions** $S_{i,j}$.

In order to create the Strong, Neutral, and Weak partitions both subject specific and subject pair impostor distributions are considered. Given two subjects, i and j, we gather S_i, S_j, and $S_{i,j}$. For S_i and S_j, we determine the maximum and minimum score - giving us a maximum and minimum value for each subject. Then, for each score in $S_{i,j}$, the difference is calculated between each subject's maximum and minimum. If a score is closer to both maximums, we partition this image pair and score as Strong - given that larger scores clearly imply the pair to be impostors. If a score is closer to both minimums, we partition this image pair and score as Weak, since these types of scores often cause false negatives. Otherwise, if the score is closer to one subject's minimum and the other subject's maximum, we partition this image pair and score as Neutral.

In the GBU challenge problem, three constraints were used to sustain balance in the partitions and remove certain causes of bias. In the following two sections we discuss the constraints considered for our experiment in relation to those used in the GBU face recognition challenge.

[1] These numbers vary from those reported in [7] due to a subject given two different subject identifications.

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



a b c

Figure 1. (a) Score distributions for the same session impostor scores and the different session impostor scores. (b) Cumulative Density Function of the two score distributions. (c) The similarity score cutoff value plotted against the corresponding False Accept Rate for the two score distributions paired with all different day authentic scores.

2.1. Same Day vs. Different Day Impostor Scores

One of the constraints imposed when selecting the GBU partitions is that the images in all authentic pairs considered were taken on different days. This is because it has been shown that images taken on the same day consistently have a higher recognition rate than images taken on different days [5]. Through our initial investigation of the data using FRVT Fusion scores used in [7], we have found this to be the case for impostor scores as well. That is, impostor image pairs taken on the same day in the same location tend to have scores which lie closer to the authentic distribution.

During image acquisition, images were taken at the same location for a series of three days, and the location then potentially changed for acquisition the following week. Subjects were expected to participate only once per week. A subject's participation is referred to as a session. Hence, when considering impostor scores, we must extend the different day constraint to a different session constraint. Figure 1.a compares the impostor score distribution from same session comparisons to the impostor score distributions from different session comparisons. For these scores, large negative scores more strongly imply an authentic score, whereas scores closer to and larger than zero more strongly imply an impostor score. A clear shift to the left towards scores which imply authentic scores is seen in the same-day impostor scores. A Kolmogorov-Smirnov test shows that the two distributions are in fact statistically significantly different, with a p-value less than $2.16 * 10^{-16}$. Figure 1.b shows the cumulative density functions used to determine the ks-statistic, and in turn the p-value, for the same session and different session impostor distributions. Additionally, if we pair the different day authentic scores from the Fusion score matrix, we can compare similarity score cutoff values to their corresponding false accept rates, shown in Figure 1.c. The score for any given FAR is more clearly an impostor for different-session impostor pairs then for same-session impostor pairs. Thus we enforce a different-session constraint in selecting data for our experiment, leaving us with 83,000,234 impostor scores.

2.2. Impostor Bias from Distinct Images and Balanced Subject Counts

Two additional constraints were imposed in the formation of the GBU problem [7]. The first constraint stated that an image can only be present in one target or query set. Within the authentic distributions, scores are computed between one image and all other images of the same subject, which is a relatively small number of comparisons overall. However, in the impostor distribution, scores are computed by comparing one image to all other images not of that subject, which is a large number of comparisons from the overall distribution. Figures 2 and 3 provides an example of the same image which contributes to image pairs whose scores have been partitioned as Strong, Neutral, and Weak using the FRVT Fusion algorithm used in the analysis of the GBU partitions [7]. This example illustrates the challenge faced when attempting to use the distinct images constraint. Namely, if we were to enforce a distinct-image constraint, how would we choose which partition would be given an image pair using this gallery image? It is not only that it would be hard to decide where to put an image pair involving the top image, it is that one of the main goals of the challenge is to understand what makes that top image pair with the different lower images to give different levels of difficulty. Hence, we do not employ a distinct image constraint in our experiment.

The second constraint stated that the number of images per person must be the same in all target and query sets.

Gallery

Strong Probe

Neutral Probe

Weak Probe

Figure 2. Example of Gallery Image (04221d582) with probe images (when paired) which fall in the Strong (05210d240), Neutral (05120d257), and Weak (05210d183) Partition. This illustrates why the distinct-images constraint and the balanced-subject-count constraint used in the GBU problem are not used in the SNoW problem; see the text for details. Additionally, this example emphasizes that there is not a single covariate, such as illumination, expression, gender, or ethnicity, which determines the partition.

When considering authentic scores, this constraint is easily met since each image pair contains images of the same subject. Thus when partitioning on authentic scores, an image pair contributes an image of the same person to both a target and a query set. However, when partitioning the impostor distribution an image pair contains images of two different subjects. Thus, in order to balance the subject count within a partition, a second image pair using the same subjects which also falls in the same partition would need to be found. This type of image pair symmetry is not guaranteed, and thus, the balanced subject count constraint is also not used for our experiment. Consider again the example in Figure 2. In order to have a balanced subject count, an image pair which contains these images would need to be found for all three partitions. Although it is likely that at some point during partitioning the partitions will contain balanced subject count, it is difficult to employ this constraint without penalizing one or more partitions.

3. Performance

In order to analyze the performance implications of each partition, three algorithms and their fusion were used. The three algorithms, listed here as AA, BB, and CC, were three of the top performers in the FRVT 2006 evaluation [8]. The resulting scores of these three algorithms were then fused as described in [7] yielding a "Fusion" score. Since the distinct images and balanced subject count constraints were removed for this problem, all different day authentic scores were used when computing score distributions and during ROC analysis for all three partitions.

Algorithm	Strong	Neutral	Weak
AA	79,301,534	1,589,280	2,109,420
BB	81,732,610	797,602	470,022
CC	79,972,582	1,989,516	1,038,136
Fusion	77,760,142	3,295,336	1,944,756

Table 1. Number of image pairs in each partition per algorithm.

Table 1 list the number of image pairs in each partition for each algorithm. In all cases, the Strong partition contains the majority of image pairs. Further, for all algorithms excluding AA, the Neutral partition contains the second largest number of image pairs, followed by the Weak partition with the fewest image pairs. This is a desirable affect for any algorithm since this implies that most scores are clearly impostors, and many of the remaining images, those in the Neutral partition, are also probable impostors. Thus a relatively small percentage of images are likely to cause false accepts. Additionally, these values reflect the effects

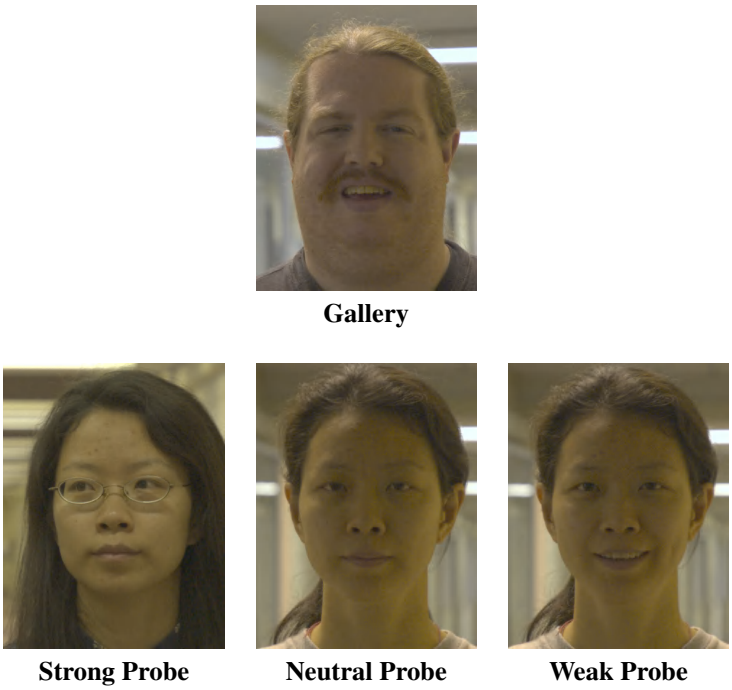


Figure 3. Example of Gallery Image (04233d529) with probe images (when paired) which fall in the Strong (04202d595), Neutral (04202d583), and Weak (04202d583) Partition. This example illustrates how the strength of an impostor score can appear to be independent of three common face convariates - gender, age, and ethnicity

of the fusion technique. With respect to the three individual algorithms, the Fusion algorithm produces fewer Strong image pairs and many more Neutral pairs. The Fusion algorithm also finds a larger than average number of Weak image pairs.

Figure 3 displays the score distribution of each partition with respect to the authentic distribution. For each algorithm a clear shift towards the authentic distribution is seen as we move from Strong to Weak. The Weak partition has a larger percentage overlap with the authentic distribution, likely the cause of most false accepts. Further, each partition overlaps the other two partitions to some extent. This implies that there is not a consensus between all image pairs about the difficulty of impostor score comparison between them.

Additionally, the legend on each algorithm's distribution graph in Figure 3 reports the bounds on the scores found in each partition. This further shows how the partitions are overlapping. For instance, the score range of the Neutral partition in algorithm AA is almost completely contained in the corresponding Strong partition. However, the mean of these partitions is very different. A similar phenomenon occurs for algorithm CC and is reflected in the Fusion algorithm, where the entire Neutral partition score range is contained in the Strong partition score range.

ROC curves were then generated for each partition using all different day authentic scores per algorithm, and are shown in Figure 3. In all four cases, the Strong partition provides the best performance, followed by the Neutral partition, and lastly the Weak partition. The large amount of overlap in the Weak partition's distribution and the authentic distribution is reflected most prominently in the performance of algorithms AA and CC. Further, Table 2 provides the True Accept Rate for each partition per algorithm based on a False Accept Rate of 0.001 computed using global thresholding. The performance of AA and CC appears much poorer than that of BB or the Fusion method. This is due to the extreme overlap between the partitions and the authentic distribution seen in the score distributions. With a better understanding of the algorithms it may be possible to select an improved point of reference for the partitioning.

Algorithm	Strong	Neutral	Weak
AA	0.182	0.000	0.000
BB	0.858	0.668	0.323
CC	0.018	0.000	0.000
Fusion	0.925	0.767	0.461

Table 2. True Accept Rate (TAR) at a False Accept Rate (FAR) of 0.001 for each partition per algorithm.

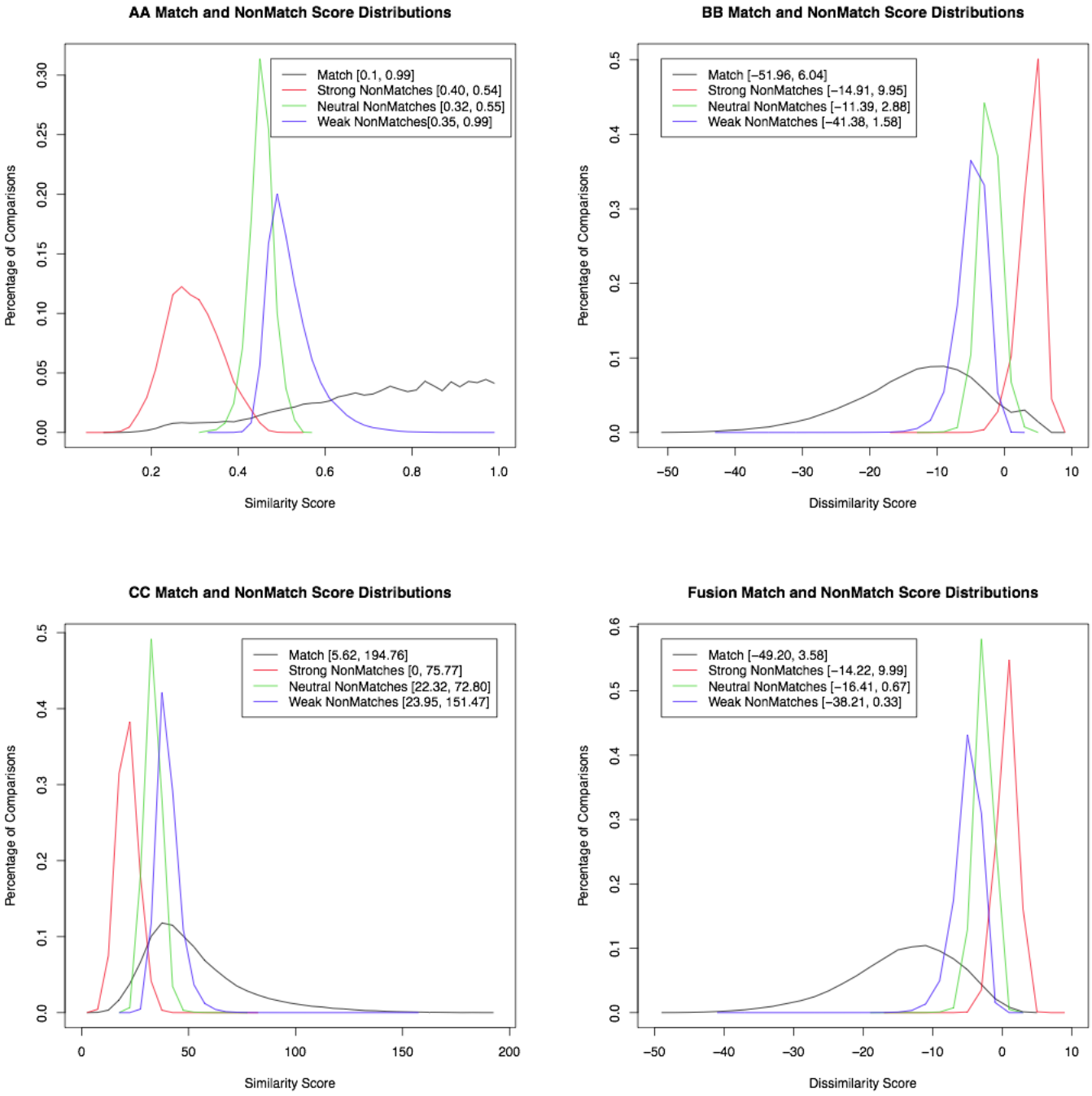


Figure 4. (Top Left) Authentic and Impostor Partition Distributions for Algorithm AA. (Top Right) Authentic and Impostor Partition Distributions for Algorithm BB. (Bottom Left) Authentic and Impostor Partition Distributions for Algorithm CC. (Bottom Right) Authentic and Impostor Partition Distributions for the Fusion Scores of Algorithms AA, BB, and CC.

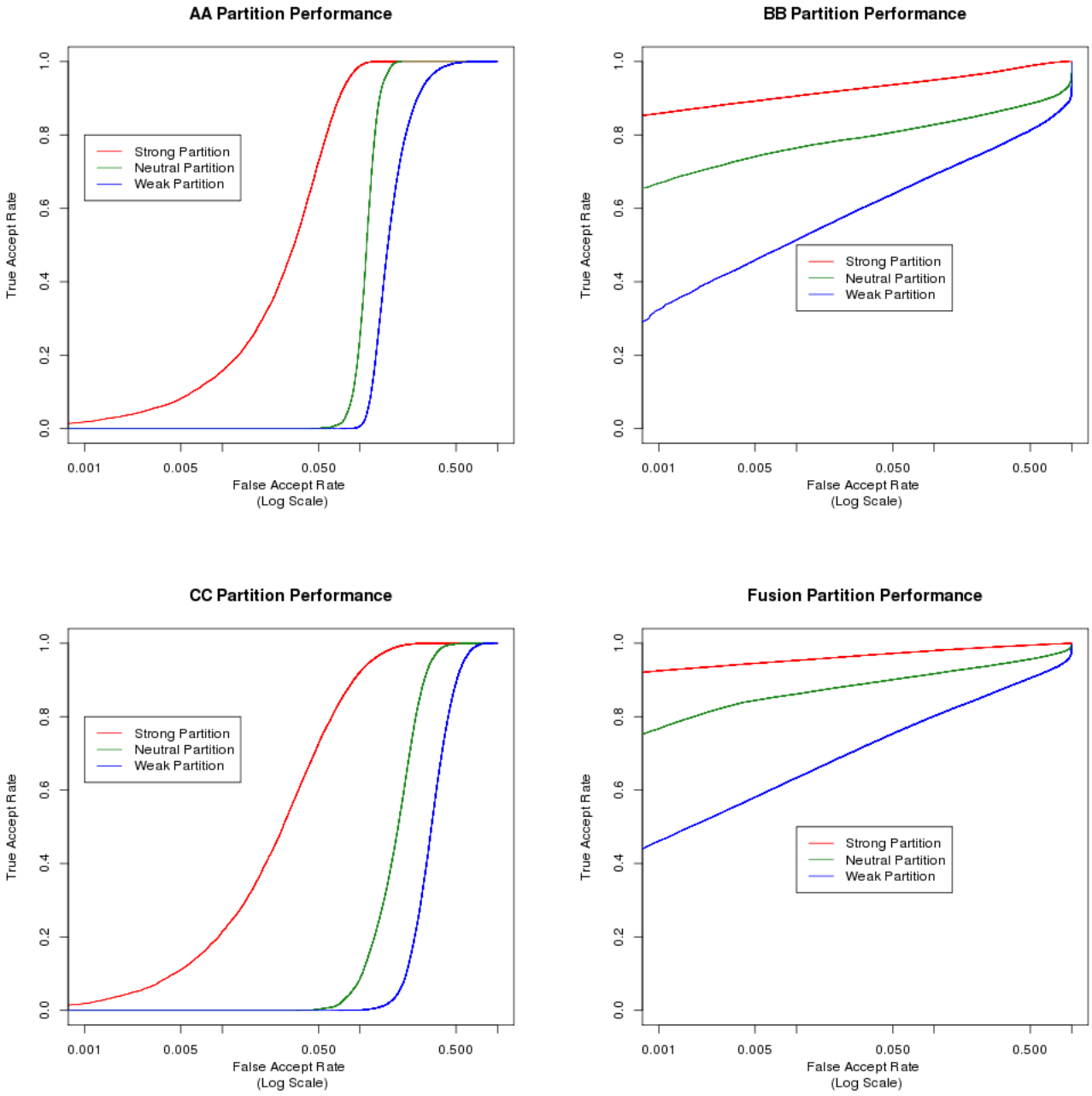


Figure 5. (Top Left) ROC performance results for the partitions developed from Algorithm AA. (Top Right) ROC performance results for the partitions developed from Algorithm BB. (Bottom Left) ROC performance results for the partitions developed from Algorithm CC. (Bottom Right) ROC performance results for the partitions developed for the Fusion Scores of Algorithms AA, BB, and CC.

4. Discussion and Conclusions

From this experiment we have shown that there are varying degrees of impostor scores represented by the Strong, Neutral, and Weak partitions. The Strong partition contains the most desirable type of impostor score, and also represents the majority of impostor image pairs. However, by comparing the local and global distributions defined by an image pair we can define a partition of scores which are likely to cause a false accept, the Weak partition. The partitions also have a clear effect on an algorithm's performance in a recognition scenario. When using the Strong partition's scores, we achieve improved performance over using the Neutral or Weak scores.

Additionally, this experiment has presented evidence that the impostor distribution deserves further exploration. Since degrees of difficulty can be determined in the impostor score, it is reasonable to consider an algorithm which accounts for scores that fall in the weak partition given some metric for weak scores.

A future goal of this work is to look at image pairs which fall in the same partition across algorithms. By determining these core groups of image pairs for each partition, we can highlight the structural weaknesses or training biases imposed by a particular algorithm. This may also include us investigating stronger reference points for partitioning. Similarly, we are interested in image pairs which move from one partition in Algorithms AA, BB, and CC, to a new partition in the Fusion algorithms. This can help us understand the impact of this type of fusion on the impostor distribution.

Another future goal is to explain the correlation between various face algorithms. This is in hope that we will be able to determine what makes an image pair fall in a particular partition, and thus an image pair's potential of causing a false accept.

References

[1] G. Aggarwal, S. Biswas, P. Flynn, and K. Bowyer. Predicting good, bad, and ugly match pairs. *Proc. 2012 IEEE Workshop on Applications of Computer Vision*, pages 153–160, 2012.

[2] J. Beveridge, P. Phillips, G. Givens, B. Draper, M. Teli, and D. Bolme. When high-quality face images match poorly. *Int'l Conf. on Automatic Face and Gesture Recognition*, pages 572–578, March 2011.

[3] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheeps, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. *Proceedings of ICSLD*, 1998.

[4] B. F. Klare and A. K. Jain. Face recognition: Impostor-based measures of uniqueness and quality. *Int'l Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, pages 237–244, 2012.

[5] Y. Lui, D. Bolme, B. Draper, J. Beveridge, G. Givens, and P. Phillips. A meta-analysis of face recognition covariates. *Int'l Conf. on Biometrics: Theory, Applications, and Systems (BTAS)*, pages 1–8, September 2009.

[6] A. O'Toole, P. Phillips, X. An, and J. Dunlop. Demographic affects on estimates of automatic face recognition performance. *Image and Vision Computing*, 30(3):169–170, 2012.

[7] P. Phillips, J. Beveridge, B. Draper, G. Givens, A. O'Toole, D. Bolme, J. Dunlop, Y. Lui, H. Sahibzada, and S. Weimer. An introduction to the good, the bad, and the ugly face recognition challenge problem. *Int'l Conf. on Automatic Face and Gesture Recognition*, pages 346–353, March 2011.

[8] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale results. *IEEE Trans. PAMI*, 32(5):831–846, 2010.

[9] W. J. Scheirer, A. Rocha, R. J. Micheals, and T. E. Boult. Meta-recognition: The theory and practice of recognition score analysis. *IEEE Trans, PAMI*, 33(8):1689–1695, 2011.

[10] M. Teli, J. Beveridge, P. Phillips, G. Givens, D. Bolme, and B. Draper. Biometric zoos: Theory and experimental evidence. *Proc. 2011 International Joint Conference on Biometrics*, pages 1–8, October 2011.

[11] N. Yager and T. Dunstone. The biometric menagerie. *IEEE Trans. PAMI*, 32(2):220–230, 2010.