NISTIR 6719

Meta-Analysis of Face Recognition Algorithms

P. Jonathon Phillips
Elaine M. Newton

U. S. DEPARTMENT OF COMMERCE
Technology Administration
Mathematics and Computational Sciences Division
National Institute of Standards
and Technology
Gaithersburg, MD 20899-8910

March 7, 2001

U.S. DEPARTMENT OF COMMERCE
Donald L. Evans, Secretary

NATIONAL INSTITUTE OF STANDARDS
AND TECHNOLOGY
Dr. Karen H. Brown, Acting Director

NISTIR 6719

Meta-Analysis of Face Recognition Algorithms

P. Jonathon Phillips
Elaine M. Newton

U. S. DEPARTMENT OF COMMERCE
Technology Administration
Mathematics and Computational Sciences Division
National Institute of Standards
and Technology
Gaithersburg, MD 20899-8910

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

possibly contradictory or inconclusive studies and discover what may be collectively said about a given field. A meta-analysis can spot trends or provide more conclusive results from a series of inconclusive studies.

A second type of meta-analysis examines a field to identify potential methodological problem. There are two classic studies of this type from medicine. The first is the study by Hedges [32] that showed a bias in meta-analyses in medicine because of their tendency to not include unpublished studies. Published studies tend to show greater effectiveness of a new drug or medical regime than unpublished studies. Thus, meta-analysis that excluded unpublished studies would be biased towards showing greater effectiveness of a new drug or regime.

The second is the study by Colditz *et. al* [33] that showed a bias in results from non-randomized experiments in medicine. In a prototypical experiment, a test subject is randomly assigned to either an experimental regime or to a control regime. An experimenter monitors the test subjects. In a randomized test (also know as a double-blind test), neither the experimenters nor test subjects know which test subjects are assigned to the experimental regime and which subjects are assigned to the control group. In a non-randomized experiment, the experimenters know which subjects are assigned to the experimental regime, while the subjects remain unaware of this. Colditz *et. al* showed that non-randomized studies report a higher success rate than randomized studies.

In this paper, we apply meta-analysis to face recognition algorithms. Like the two previous examples, our analysis addresses experimental design methodology. Automatic face recognition is amenable to meta-analysis for a number of reasons. The first is that this has been a very active area of research for the last decade. Second, there exists an accepted quantitative performance measure -- probability of identification. For example, it would be difficult to perform a meta-analysis of edge detection algorithms where there is not an accepted measure of performance. Third, there exist databases of facial images that are available to researchers and used to report results in the literature. Fourth, there exist independent measures of performance -- the FERET evaluations. Fifth, there exists a generally accepted baseline algorithm that is easily implemented by face recognition researchers -- eigenfaces [34].

Not all five elements are required to perform a meta-analysis, but it certainly helps. In fact, there are elements missing that would enable a more detailed meta-analysis. Some of the elements missing include, statistical levels of significance for performance results, and general availability of standard evaluation protocols.

In computer vision, randomized and non-randomized experiments do not exist as defined in the arenas of medical and social scientific research, but we will define analogues. We divided face recognition experiments into three classes. The

first is the most stringent type of experiment and roughly corresponds to a double-blind test. This class consists of independently administered evaluations with sequestered data. Examples of this class are the FERET evaluations [2,3]. We use results from the FERET evaluation as the standard for evaluating the difficulty of experimental results in the meta-analysis. The second class roughly corresponds to a non-randomized experiment, and consists of experiments where the performance of a new algorithm is compared with the performance of a baseline or control algorithm. The last class consists of experiments where the performance of a new algorithm is not compared with a control or baseline algorithm.

In this paper, we will address the importance of choosing the correct evaluation methodology for conducting experiments; the role of a baseline or control algorithm in experiments; and the need to document experimental parameters, design decision, and performance results.

Methodology for Selecting Papers

In this section we will describe the method used to select the papers in the analysis presented in this paper. Over the last decade numerous papers have appeared in the face recognition literature, with virtually all of the papers presenting experimental results. The basic requirements for selecting papers for this study were that experiments used either the FERET or ORL database, and reported identification performance results for full frontal facial image.

We searched major computer vision and face recognition conference proceedings, journals, edited books, and the IEEEXplore journal and conference archive. This produced 47 papers. We sorted through these papers and removed papers that had similar experimental results from the same research group. This situation usually arose from conference and journal papers describing the same algorithm. In all but one case, when results appeared in conference and journal papers, the journal paper was selected. (In the one case, the journal paper did not contain an experimental result that met our selection criteria.) We also removed one paper from our analysis set because we could not determine the experimental performance results. The sorting process produced 24 papers for further analysis, and a list of these papers is in the reference section [4-27]. Some papers reported results for more than one algorithm, and some reported results on more than one dataset. This produced 68 performance scores that we use in our analysis.

We consolidated the results of three sets of papers. The first set of consolidated papers reported results on the ORL database using the same basic evaluation protocol [8, 9, 10, 14, 19, 22, 23]. Two of these papers also reported results on

the FERET database [10, 23]. The second set, were two papers by Moghaddam and Pentland [15, 16] that used the same image sets. The third set consisted of three papers by Liu and Wechesler [11, 12, 13] that used the same image sets.

Statistics Selected

The goal of each of the papers selected was to present a new algorithm for automatic face recognition. which will we refer to as an experimental algorithm. The ability of the experimental algorithm was demonstrated by experiment(s) in each paper. To analyze these experimental results, it was necessary to extract the following experimental parameters for each experiment and algorithm in our final set of papers:

1. Identification performance score
2. How scores were reported (in Text, Table, or interpreted from a Graph)
3. For graphical data, the error introduced by reading the score from the graph.
4. Are the gallery and training sets the same?
5. Number of Training Images
6. Number of People in the Training Set
7. Number of Gallery Images
8. Number of People in Gallery Set
9. Number of Probe Images
10. Number of People in Probe Set

We restricted our analysis to identification performance scores that reported the fraction of probes that were correctly identified. A *gallery* is a set of known individuals against which an algorithm attempts to perform recognition. A *probe* set is a set of images of unknown individuals that an algorithms attempts to recognize. For the identification performance used in this paper, a probe is a new image of an individual in the gallery. The identification score was selected because this is the performance measure of choice for the vast majority of papers in face recognition (item 1 above). Papers reported either accuracy or error rates. For our analysis, accuracy rates were converted to error rates. Some papers reported additional scores, which we did not include in our analysis.

Performance scores were reported in at least one of three ways in each paper: numerically in the text or a table and/or graphically (item 2). Numeric scores were selected over scores reported in a graph. If only a graph was available,

performance scores were interpreted from the graph, and subsequently a small error was introduced into our data. This error range was estimated and recorded for each graphically interpreted quantity (item 3).

In face recognition experiments, there are three sets of images: training, gallery, and probe. The *training* set is used to generate a face representation and to tune algorithm parameters. For example, in PCA-based algorithms the eigenfeatures are generated from the training set.

Performance scores are computed using a gallery and probe set. The training set and the gallery do not have to be the same set of images. When the training set and gallery are the same, algorithm performance is usually better.

From each paper, we attempted to obtain the total number of images and the number of different individuals in each of the training (items 5 and 6 above, respectively), gallery (items 7 and 8 above, respectively), and probe (items 9 and 10 above, respectively) sets. Few papers report all of this information. Some papers report training or gallery information but did not clearly report if their training and gallery sets are the same (item 4). Item four above is a record of whether the training and gallery sets are the same or if the relationship is unclear in the paper. Some papers used the terminology "test set," "images used in testing," or "query set" and not the term "probe set." If the terms were not adequately defined in the paper, these terms were assumed to have the same meaning. If the term "probe" was not used, we recorded the term that appeared to have the same meaning.

If the authors reported performance for a number of variations for an algorithm, we choose the variation that had the best overall performance. For the consolidated ORL algorithms, we selected the performance score that corresponded to the de facto ORL evaluation protocol.

A number of papers reported performance scores for additional algorithms that served as controls. If there was only one control algorithm, we refer to this as the baseline algorithm for the experiment. In this case, the baseline algorithm was usually a correlation- or PCA-based face recognition algorithm. (PCA-based face recognition is also known as eigenfaces.)

Some papers present multiple control algorithms. In this case we selected the variation of a PCA-based algorithm with the best performance as the baseline algorithm. We selected the PCA-based algorithm for uniformity in our analysis and because PCA-based algorithms are the de facto baseline standard in the face recognition community.

Analysis of Performance Scores

The first step in the analysis is to look at the distribution of the identification error rates across all experiments and algorithms -- experimental and control (control contains the baseline). Figure 1 is a histogram of these error rates, which shows 38 out of 68 (56%) have and error rate below 0.10.

Next we restrict our attention to the experimental and baseline algorithms. There are 40 experimental algorithms, and 33 experimental algorithms with corresponding baselines. There are less baseline algorithms because some baselines are used for more than one experimental algorithm; e.g., the ORL series has one baseline algorithm for 10 experimental algorithms.

Figure 2 shows a histogram of error rates for experimental algorithms in black and baseline algorithms in white. To illustrate the influence of a baseline score, we count a baseline score by the number of times that is served as a baseline. From the ORL example, we counted the baseline algorithm 10 times. Of the 40 experimental algorithms, 29 (73%) report error rates of 0.10 or less.

We first look at the seven experimental algorithms that do not have a baseline score. The error rates for these algorithms are: 0.008, 0.01, 0.02, 0.034, 0.045, 0.046, and 0.28. Their median is 0.034. These scores (1) show 6 out of 7 experiments have an error rate less than 0.05, (2) contain the best error rate (0.008) for all 40 experimental algorithms in this analysis, and (3) account for one third of the experimental algorithms with error rates below 0.05. Clearly, the results from experimental algorithms without a supporting baseline algorithm are highly biased.

The next group of algorithms that we looked at was the consolidated ORL algorithms. The error rate for the ORL baseline PCA algorithm [8] is 0.105. The error rate range for the experimental algorithm was 0.029 to 0.13, with 7 out of 10 performance scores equal to or less than
0.05. This indicates that performance has been saturated using this evaluation protocol, and the protocol does not define a sufficiently challenging problem for automatic face recognition.

Next we examine the relationship between the baseline and the experimental scores. There are 33 experimental algorithms with a baseline score, and 24 out of 33 (73%) baseline scores have an error rate less than 0.20. Of these 24 algorithms, 21 of the experimental algorithms have an error rate less than 0.10. The median performance score for the nine experimental algorithms with baseline scores greater than 0.20 is 0.31. The median performance score for the 24 experimental algorithms with baseline scores less than 0.20 is 0.05. The median performance score for the 21 experimental

algorithms with baseline scores less than 0.10 is 0.041. This shows that the performance of the baseline score can be used as a predictor of the performance of the experimental algorithms. The baseline scores in turn can serve as a benchmark for characterizing the difficulty of the experiment used to measure algorithm performance. We now proceed to specifically address the relationship between algorithm scores and characterizing the difficulty of an experiment.

In computer vision and face recognition, the difficulty of the problem solved is defined by the image sets, evaluation protocol, and scoring method used in the experiment. To be able to make a statement about the difficulty of the problem addressed by the majority of the face recognition experiments examined in our analysis, we need to compare it to a well-studied baseline. We do this by comparing the above results to the Sep96 FERET evaluation [2]. The FERET evaluation reports performance for a number of classes of problems. The two most relevant classes for this analysis are FB and dup I probe categories. In the FB probe category, algorithms are asked to recognize facial images when the gallery and probe images are taken within five minutes under the same conditions. In the dup I probe category, algorithms are asked to recognize faces when the gallery and probe image of a person are taken on different days or under different conditions on the same day.

As shown in the FERET evaluations, the FB probe category represents the easiest possible problem in face recognition, and provides an empirical upper bound on the current state of automatic face recognition, and does not represent an interesting practical problem. In the Sep96 FERET evaluations, the best face recognition algorithms achieve error rates of less 0.05 on a database of 1196 individuals with one image per person [2]. The ease of this category of problems has been known since the first and second FERET evaluations were administered in August 1994 and March 1995 [3].

The dup I category represents a very interesting and practical problem, identifying faces when the gallery and probe images are taken on different days. In the Sep96 FERET evaluation, the best error rates are between 0.40 and 0.45 on a database of 1196 individuals. This shows that the dup I problem is clearly a difficult problem.

By comparing performance of the experiments in this analysis with performance scores on FERET data, we will show that the majority of the experimental results from the papers in this study are equivalent to the FB category. We demonstrate this by comparing performance from a FERET Sep96 baseline algorithm and FERET evaluated algorithms. The baseline algorithm was a PCA-based algorithm that used the L1 metric in the nearest neighbor classifier.

A robust set of performance statistics was obtained from FERET; we computed error rates from 100 galleries of size 200. These are the same galleries used in Moon and Phillips [1]. All galleries were randomly generated with replacement from the same pool of 1196 distinct faces (note: the pool contained one facial image of 1196 individuals). However, each individual gallery was generated without replacement. Thus, there is overlap between galleries, but not within galleries, and each gallery contains facial images of 200 individuals.

All algorithm training was performed using the development portion of the FERET database, and was completed to prior to taking the Sep96 FERET evaluation. Thus, all training was completed prior to calculating performance on the 100 galleries. Some of the pool of 1196 images used to generate the 100 galleries is in the FERET development set. Thus, the training set and the galleries are not independent, but the galleries and training set are not the same. For each gallery, the FB probe set was generated by searching the FERET database for frontal images taken on the same day under the same conditions—by the design of the FERET database there exist one such image for each gallery image. The dup I probe set for each gallery was generated in a similar manner. There is not a duplicate image for all images in the pool of 1196 images. Thus, the number of duplicates probes associated with each gallery varies. For each of the 100 galleries we generated the identification error rates for the corresponding FB and dup I probe sets. Thus, for each algorithm there are 100 FB error rates and 100 dup I errors.

Figure 3 shows the 100 FB scores for the PCA baseline, MIT95 and MIT96, MSU, UMD97, and USC algorithms. We will call the MIT95 and MIT96, MSU, UMD97, and USC algorithms collectively the "FERET evaluated algorithms." The scores are shown on a box graph. On a box graph, the line in the center of the box represents the median, the two horizontal edges are the quartiles, the whiskers represent the range of the data that is within 1.5 times the quartile distance, and dots represent outliers. Figure 3 shows that all of the baseline scores are less than 0.20 and that vast majority of the FERET evaluated algorithms scores are less than 0.10. This corresponds with the results observed in analyzing the performance scores of the experimental algorithms with a baseline score of less than 0.20.

Figure 4 shows a boxplot of the 100 dup I scores for the same algorithms in figure 3. This shows a much greater range of scores for both the baseline scores and FERET evaluated algorithms. Once again, this corresponds to the experimental algorithms in this analysis with corresponding baseline scores greater than 0.20.

By comparing the relative performance of the baseline and experimental from the experiments in the meta-analysis paper suite and the FERET results we have shown that the majority of experiments in face recognition papers have

concentrated on a relatively easy task. By reporting performance on datasets that saturate performance levels, it is hard to demonstrate significant breakthroughs in automatic face recognition. The papers by Teh and Hinton [25] and Bartlett, Lades, and Sejnowski [4] are nice examples of testing done on both FB and dup I equivalent problems.

Experimental Parameters

One of the key sets of information for interpreting, understanding, and independently reproducing experimental results are the experimental parameters. When tallying the number of papers that presented particular types of information clearly, we evaluated the parameters of all of the experiments included in this analysis. If a paper from which we reported multiple experiments was unclear in reporting parameters for one or more experiments, it was counted as being unclear. For face recognition experiments, the key experimental parameters are the performance results and selection criteria for training, gallery and probe sets. The selection criteria are parameterized by the statistics discussed in the section on "Statistics Selected." Knowledge of the contents and relationship among these three sets in critical to evaluating the results of an experiment. When we attempted to obtain the total number of images and the number of different individuals in each of the training, gallery, and probe sets, we discovered that this information was not always reported or was reported in an ambiguous way. Only twelve of the 24 papers reported all of this information in a clear manner for their experimental and baseline algorithm tests, which amounts to 22 of the 40 corresponding performance results.

An often-overlooked design decision for experiments is the relationship between the training set and gallery. When they are the same, performance is usually better. Eighteen papers report training or gallery information but did not clearly report if their training and gallery sets are the same. Sixteen out of 24 papers clearly defined the relationship between the gallery and training set, which represents 30 out of 40 performance results.

The goal of including experiments in a paper is to inform the reader of the outcome(s) of the experiments. The reader is informed by the performance results in the paper, and the results should be presented so that the reader can easily ascertain them. Out of the 24 papers that we used in this analysis, 15 reported all of their performance scores clearly in a table or in the text, which account for 22 out of 40 experimental performance results and their corresponding baseline scores (if available). Of the nine papers that did not report performance scores in a clear manner, one was unclear because the authors did not disclose their definition of a positive identification result, which can be defined in several ways depending on how many images the algorithm retrieves as a possible match. This paper accounts for two of the

performance score sets. (All of the results we use in our analysis are for the algorithm's performance in finding a match in the first retrieved image.) The remainder of the papers presented either or both of the experimental and baseline algorithm performance scores only in a graph.

Interpreting a point from a line or bar graph can present error in several ways. Sixteen out of the 40 sets of experimental and baseline performance scores in this analysis are subject to error introduced by graphical interpretation. (One of the 16 experimental scores did not have an accompanying baseline score.) These sets of data were taken from 11 different graphs. (The number of graphs and the number of data sets are not equal because data was presented in different combinations in each paper. For example, some papers showed multiple experimental and baseline algorithms' performance scores all on one graph. Others showed this information in separate graphs.) Six graphs had a performance score axis divided into intervals of 0.05 or 5%; two were in intervals of 0.10 or 10%; two were in intervals of 0.02 or 2%; and one was in intervals of .20 or 20%. Only two of the graphs had lines across the graph at each major unit interval for performance scores. None of these graphs provided minor unit lines across the graph or tick marks along the axis showing the performance score. To obtain a performance score, one must first correctly pick the point on the graph that corresponds to the desired data point and decide what fraction correctly describes the distance between where the point falls on the axis and the two closest major interval units on the axis. The larger the interval between unit marks on the axis, the larger the error introduced when reading a score off of the graph. This is compounded by small graph sizes, which may have been distorted in the reproduction process.

This analysis of experimental parameters shows that authors need to spend more effort in describing the details of their experiments, and reviewers need to carefully read papers to insure that all the necessary parameters are included. Nearly half of the papers fall short of demonstrating their main objective (to present a complete case for a new face recognition) by not completely quantifying the performance of their algorithm. The inclusion of this information is necessary if authors are to make a compelling case for the accomplishments of their research.

Conclusions

In this meta-analysis, we have identified two areas of methodological problems in the way face recognition experiments are performed and presented. The first area of problems is the type of experiments performed. The majority of experiments in face recognition have concentrated on experiments that have error rates less than 0.10 and subsequently

saturate the performance level. Researchers need to concentrate on face recognition problems that are harder, as defined by the image sets in the experiments and the performance by a baseline algorithm. One reason that researchers report very low error rates is to convince other researchers that their algorithm performs well and thus makes a scientific contribution to automatic face recognition. When available, independent evaluations are gold standards for establishing the contribution of a new face recognition algorithm. In the absence of an independent evaluation, the performance of baseline algorithms can serve as a yardstick for measuring the contribution of a new algorithm, as demonstrated in this paper by experimental performance scores that have corresponding baseline scores which fall above and below 0.20. The best baseline algorithm would be a standard implementation of a face recognition algorithm that is readily available to all researchers. To establish a sound foundation for the incorporation of standard baseline algorithms into an experimental method, it is necessary to establish accompanying standard evaluation protocols and image sets. This would allow researchers to assess performance of new algorithms using established methodologies. This idea is similar to that pursued by the ORL sequence of experiments. This would be an improvement over ORL because (1) algorithms would be evaluated on a harder face recognition problem, and (2) the performance scores would be generated from exactly the same partition of a data set, not on similar partitions.

The second problem area is how authors describe the experiments and present the results. More attention needs to be paid to the details of the experimental design by authors, reviewers, and editors. Attention to these details will make for papers that are more readable and will allow researchers to fully evaluate the contribution of a new algorithm and allow for independent replication of published performance results.

Correcting the methodological issues raised in this paper will help put the development and analysis of automatic face recognition algorithms on a solid scientific basis and improve the quality of discourse in the field. This will also serve as an example for other areas of computer vision.

References

[1] H. Moon and P. J. Phillips, "Computational and performance aspects of PCA-based face recognition algorithms," *Perception*, (in press).

[2] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET Evaluation methodology for face-recognition algorithms," *IEEE trans. PAMI*, Vol. 22, No. 10, 2000.

WHERE IS PAGE 12?

[16]B. Moghaddam and A. Pentland, "Beyond Linear Eigenspaces: Bayesian Matching for Face Recognition," *Face Recognition: From Theory to Applications* (H. Wechsler, P.J. Phillips, V. Bruce, F.F. Soulié, and T.S. Huang, eds.), pp. 230-243, Berlin: Springer-Verlag, 1998.

[17]B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Presentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp. 696-710, July 1997.

[18]K. Okada, J. Steffens, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. von der Malsberg, "The Bochum/USC Face Recognition System and How it Fared in the FERET Phase III Test," *Face Recognition: From Theory to Applications* (H. Wechsler, P.J. Phillips, V. Bruce, F.F. Soulié, and T.S. Huang, eds.), pp. 186-205, Berlin: Springer-Verlag, 1998.

[19]L. Pessoa and A.P. Leitão, "Complex Cell Prototype Representation for Face Recognition," *IEEE Transactions on Neural Networks*, Vol. 10, No. 6, pp. 1528-1531, November 1999.

[20]P.J. Phillips, "Matching Pursuit Filters Applied to Face Identification," *IEEE Transactions on Image Processing*, Vol. 7, No. 8, pp. 1150-1164, August 1998.

[21]P. J. Phillips, "Support Vector Machines Applied to Face Recognition," Technical Report NISTIR 6241, *Advances in Neural Information Processing Systems 11*, eds. M.J. Kearns, S.A. Solla, and D.A. Cohn, MIT Press, 1999.

[22]F.S. Samaria, "Face recognition using hidden Markov models," Ph.D. dissertation, Trinity College, Univ. Cambridge, Cambridge, U.K., 1994.

[23]T. Sim, R. Sukthankar, M. Mullin, and S. Baluja, "High-performance memory-based face recognition for visitor identification," Technical Report JPRC-TR-1999-001-1, Just Research, 1999.

[24]D. L. Swets and J. Weng, "Discriminant Analysis and Eigenspace Partition Tree for Face and Object Recognition from Views," *Proceedings, Second International Conference on Automatic Face and Gesture Recognition*, pp. 192-197, 1996.

[25]Y.W. Teh and G. E. Hinton, "Rate-coded Restricted Boltzmann Machines for Face Recognition," Neural Information Processing Systems, 2000.

[26]L. Wiskott, J. Fellous, N. Krüger, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp. 775-779, July 1997.

[27]W. Zhao, R. Chellappa, and P.J. Phillips, "Subspace Linear Discriminant Analysis for Face Recognition," Technical Report, 1999.

[28]L. V. Hedges and I. Olkin, *Statistical methods for meta-analysis*, Academic, New York, 1985.

[29]R. Rosenthal, *Meta-analytic procedures for social research (revised)*, Sage: Beverly Hills, CA 1991.

[30] J. P. Greene, "A meta-analysis of the Rossell and Baker review of bilingual education research," *Bilingual Research Journal,* Vol. 21, No. 2 & 2, 1997.

[31] P N Shapario and S D Penrod, "Meta-analysis of face identification studies," *Psycological Bulletin,* Vol. 100, pp 139-156, 1986.

[32] L. V. Heggs, "Modeling Publication Selection Effects in Meta-Analysis," *Statistical Science,* Vol. 7, No. 2. pp 246-255, 1992.

[33] G. A. Colditz, A. Miller, F. Mosteller, "How study design affects outcomes in comparisons of therapy. I: medical," *Statistics in Medicine*, Vol. 8, pp 441-454, 1989.

[34] M. Turk, A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience,* Vol. 3, No. 1, pp 71-86, 1991.

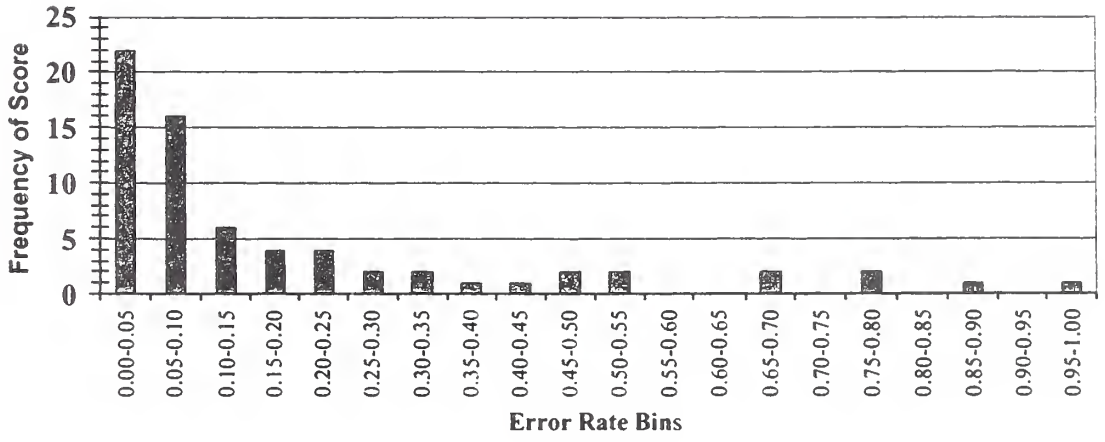

Figure 1. Histogram of all experimental and control (including baseline) error rates scores. Each bar corresponds to scores greater than the lower interval and equal to or less than the upper interval.

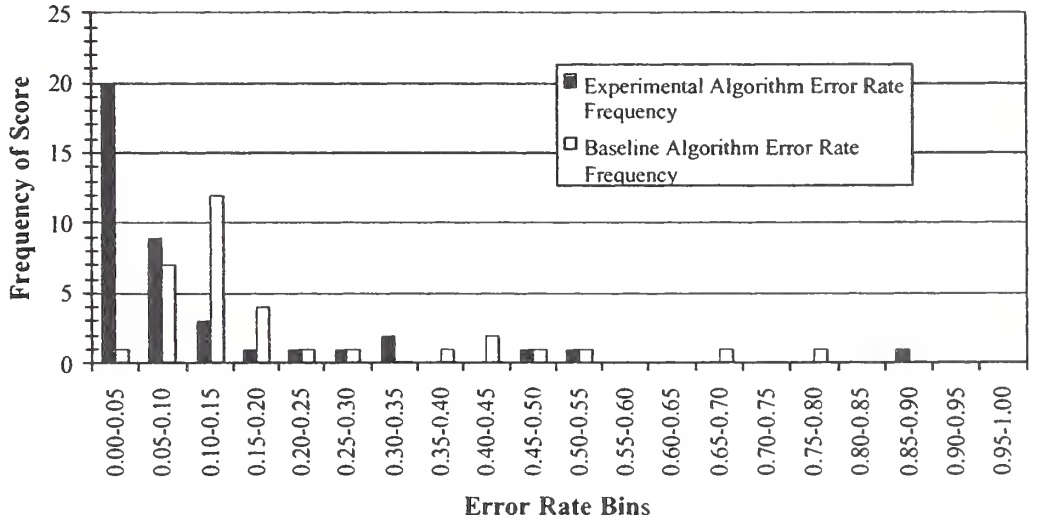

Figure 2. Histogram of error rates for experimental and baseline algorithms. Each bar corresponds to scores greater than the lower interval and equal to or less than the upper interval.

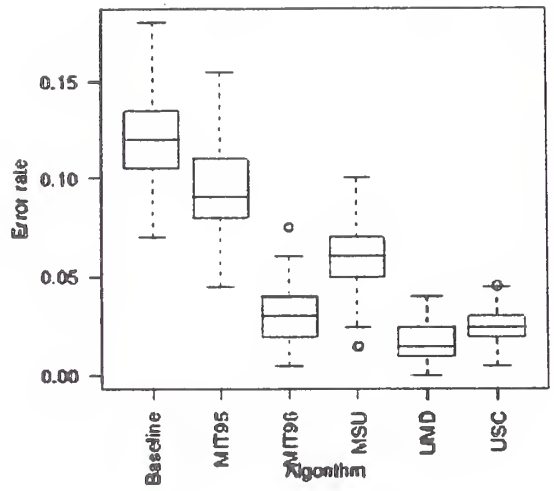

Figure 3. Boxplot of FB scores for 100 galleries for the PCA baseline, MIT95, MIT96, MSU, UMD, and USC algorithms.

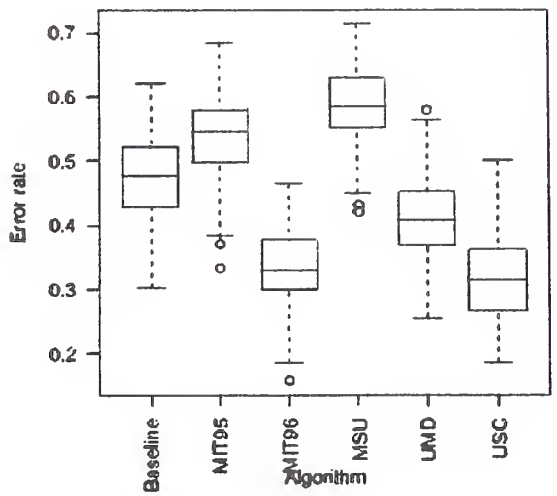

Figure 4. Boxplot of dup I scores for 100 galleries for the PCA baseline, MIT95, MIT96, MSU, UMD, and USC algorithms.