Meta-Analysis of Face Recognition Algorithms

P. Jonathon Phillips
NIST
100 Bureau Dr., STOP 8940
Gaithersburg, Md. 20899-8940
jonathon@nist.gov

Elaine M. Newton
RAND
201 N. Craig St., Suite 102
Pittsburgh, Pa. 15213
enewton@rand.org

Abstract

To obtain a quantitative assessment of the state of automatic face recognition, we performed a meta-analysis of performance results of face recognition algorithms in the literature. The analysis was conducted on 24 papers that report identification performance on frontal facial images and used either the FERET or ORL database in their experiments. The analysis shows that control scores are predictive of performance of novel algorithms at statistically significant levels. The analysis identified three methodological areas for improvement in automatic face recognition. First, the majority of papers report experimental results for face recognition problems that are already solved. Second, authors do not adequately document their experiments. Third, performance results for novel or experimental algorithms need to be accompanied by control algorithm performance scores.

1. Introduction

Meta-analysis is a quantitative method for analyzing the results from multiple papers on the same subject [28,29]. Meta-analysis can be performed to consolidate a group of experimental results or to gain deeper insight into methodological techniques in a field. Meta-analysis has been used extensively in medicine, psychology, and the social sciences.

One type of meta-analysis is a statistical analysis of results from multiple papers on a subject from different research groups. The goal is to take the results of a number of possibly contradictory or inconclusive studies and discover what may be collectively said about a given field. This analysis can provide conclusive results from a series of inconclusive studies or spot trends that cannot be detected from a single experiment. Examples of this are the efficaciousness of Taxol for breast cancer [29], the effectiveness of bilingual education [30], and an assessment of human identification studies in psychology [31].

A second type of meta-analysis examines a field to identify potential methodological problems. Each field has its established conventions of conducting and reporting research results. It is possible that the established conventions have adverse effects on the field or skew results. In this paper, we examine the current methods for conducting and reporting results for automatic face recognition algorithms.

There are two classic studies from medicine that illustrate this category of meta-analysis. 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-analyses 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 assigned to either an experimental regime or to a control regime. In a randomized test, subjects are randomly placed in either a treatment (experimental) group or a control group. Colditz *et al.* showed that non-randomized studies report a higher success rate than randomized studies.

Like the two previous examples, our analysis addresses experimental methodological issues and conventions for face recognition algorithms. By performing a meta-analysis, not only can we quantitatively investigate the validity of the reported results, we can also report on the underlying causes and recommend possible solutions.

While the computer vision community has discussed some of the results of this analysis at the philosophical level, none have been studied quantitatively. There is a quip in the face recognition community that researchers always report algorithm performance of 95% and higher (correct identification). At the same time, independent evaluations performed by FERET [2, 3] and the Face Recognition Vendor Test (FRVT) 2000 [35]

Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02)
0-7695-1602-5/02 $17.00 © 2002 IEEE

show such performance for only one case: images taken on the same day under the same lighting conditions.

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

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, and so there is a sizeable amount of accumulated work in the area. Second, there exists an accepted quantitative performance measure—probability of identification. Third, there exist databases of facial images that are available to researchers and are used to report results in the literature. Fourth, there exist independent measures of performance, the FERET evaluations for example. Fifth, there exists a generally accepted control algorithm that is easily implemented—principal component analysis (PCA)-based algorithms (also known as eigenfaces) [34].

2. Methodology for Selecting Papers

We selected papers for this study that ran experiments using either the FERET or ORL databases and reported identification performance results for full frontal facial images.

We searched major computer vision and face recognition conference proceedings, journals, edited books, and the IEEEXplore journal and conference archive, which produced 47 papers. We then removed papers that had similar experimental results from the same research group. The sorting process produced 24 papers for further analysis. 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 data set. This produced 68 performance scores that we used in our analysis.

3. Selected Statistics

Each of the papers selected presented a new face recognition algorithm, which we will refer to as an experimental algorithm. To analyze the experimental results for these algorithms, it was necessary to extract the following experimental parameters for each experiment and algorithm in our final set of papers (descriptions of each of these parameters follow the list):

1. Identification performance score
2. How scores were reported (in the text, in a table, or interpreted from a graph)
3. For graphical data, an estimate of 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 the gallery set
9. Number of probe images
10. Number of people in the probe set

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

Performance scores were reported: numerically in the text or in 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 (as was the error, item 3).

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, respectively), gallery (items 7 and 8, respectively), and probe (items 9 and 10, respectively) sets. Few papers report all of this information. Some papers report training or gallery information but do not clearly report if the training and gallery sets are the same (item 4). Item four is a record of whether the training and gallery sets are the same or if the relationship is unclear in the paper.

If the authors reported performance for a number of variations for an algorithm, we choose the variation with 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 control algorithm for the experiment. In this case, the control algorithm was usually a correlation- or PCA-based face recognition algorithm. If there were multiple control algorithms, we selected the variation of a PCA-based algorithm with the best performance as the control algorithm.

4. Analysis of Performance Scores

4.1 Viewing the Data Through Histograms

We first looked at the distribution of the identification error rates across all experiments and algorithms (experimental and control algorithms). Of all of the error rates in this analysis, 56% (38 out of 68) have an error rate below 0.10.

Next we restricted our attention to the experimental and control algorithms according to the exclusion criteria described at the end of the previous section (Selected Statistics). This yielded 40 experimental algorithms, 33 of which have corresponding control algorithms. There are fewer control algorithms because seven studies did not use a control. Some control algorithms correspond to more than one experimental algorithm (e.g., the ORL series has one control algorithm for 10 experimental algorithms).

Figure 1 shows a histogram of error rates for experimental algorithms in black and control algorithms in white. To illustrate the influence of a control score, we counted them each time a score served as a control (for a total of 33 controls). For example, for the ORL experiments, we counted the control algorithm 10 times. Figure 1 shows that 29 of the 40 (73%) experimental algorithms report error rates of 0.10 or less.

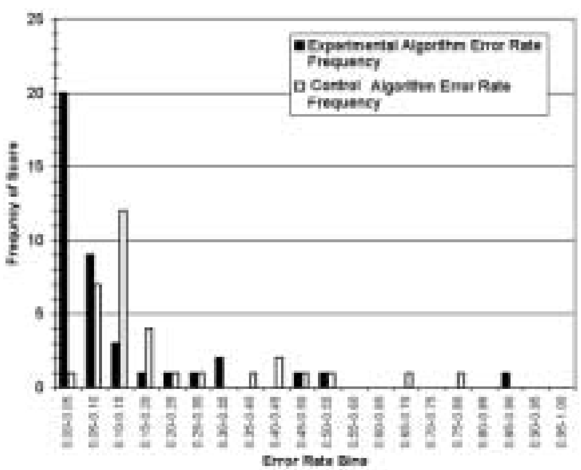

Figure 1. Histogram of Error Rates for Experimental and Control Algorithms.

We examined the seven experimental algorithms that do not have a control 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 control algorithm are highly biased.

Next we looked at was the consolidated ORL algorithms, which used the same data set and control algorithm. The error rate for the ORL control algorithm (PCA) [8] is 0.105. The error rate range for the experimental algorithms is between 0.029 and 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 data set, and the data set does not define a sufficiently challenging problem for automatic face recognition.

4.2 Difficulty of the Experiment

In this paper, we have taken an operational definition of the difficulty of a problem: how well does an algorithm or a collection of algorithms perform on a set of images. To establish an initial measure of the difficulty of a problem, we use the results from the September 1996 FERET evaluation [2]. Results are reported for independently implemented algorithms. The control algorithm was a PCA-based algorithm that used the L1 metric in the nearest neighbor classifier. 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 of one another under the same lighting 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 performance. It does not represent an interesting problem. In the September 1996 FERET evaluations, the error rates for FB probes ranged from 0.05 to 0.23. The scores are from a gallery of 1196 individuals, with one image per person [2]. The control algorithm error rate was 0.21, and the three best algorithms had error rates of 0.05. 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 September 1996 FERET evaluation, the error rates ranged from 0.41 to 0.69, and the control algorithm error

Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02)
0-7695-1602-5/02 $17.00 © 2002 IEEE

rate was 0.59. Performance was computed on the same gallery as the FB probe set.

The FERET evaluations allow for a division of frontal face recognition into two problem classes. An “easy” problem is equivalent to gallery and probe images being taken on the same day; a challenging problem is equivalent to gallery and probe set taken on different days. We propose that a problem be classified as either easy if the control algorithm error rate is below 0.20 or challenging if the control error rate is above 0.20. An “easy” problem is equivalent to gallery and probe images being taken on the same day; a challenging problem is equivalent to gallery and probe set taken on different days. The choice of 0.20 is based on analysis in this paper and additional analysis (but space constraints prohibit inclusion), as well as experience with face recognition. Better methods for categorizing face recognition problems will follow as progress is made in understanding what factors affect the difficulty of face recognition problems.

4.3 Evaluation of Experiments with a Control

Next, we examined the relationship between the control and the experimental scores from our literature search. There are 33 experimental algorithms with a control score, and 24 out of the 33 (73%) control scores have an error rate less than 0.20. Of these 24 algorithms, 21 of the experimental algorithms have an error rate of less than 0.10. The median performance score for the nine experimental algorithms with control scores greater than 0.20 is 0.31. The median performance score for the 24 experimental algorithms with control scores less than 0.20 is 0.05. The median performance score for the 21 experimental algorithms with control scores less than 0.10 is 0.041.

Table 1 shows median error rates for all scores, experimental scores with controls less than 0.20, and experimental scores less than 0.10 and with controls less than 0.20, each along with the percentage of experimental algorithms that is represented in each group.

Table 1. Median Scores for Different Sets of Experimental Scores.

	No. of Images (%)	Median Error Rate
All experimental algorithms	33 (100%)	0.07
Experimental. score with control ≤0.20	24 (73%)	0.05
Control ≤0.20 and experimental. ≤0.10	21 (64%)	0.041

Figure 2 is a scatter plot of the 33 experimental scores that had corresponding control scores. The x-axis is the experimental score, and the y-axis is the corresponding control score. A best-fit line for the data is shown in figure 2 as well. We computed the correlation coefficient value r for the 33 scores. The correlation value r is 0.932, which has a significance level greater

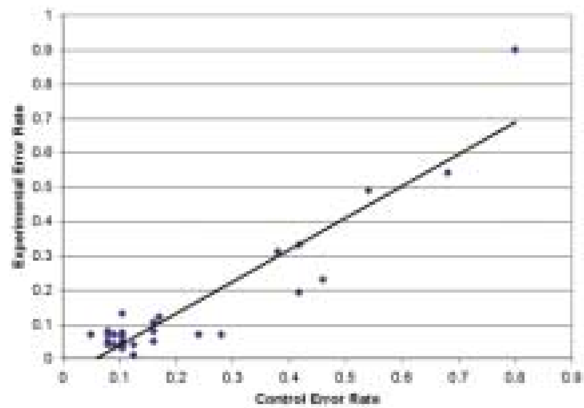

Figure 2. Best Curve Fit for Scatter Plot of Experimental vs. Control Error Rates for 33 Pairs of Scores.

than the 0.01. This shows strong correlation between the 33 pairs of control and experimental scores. Then we divided the pairs of scores into two groups and examined the relationship between experimental and control scores for pairs with control scores above 0.20 and below 0.20. Figure 3 is a scatter plot of the nine experimental algorithms with control scores above 0.20. The correlation coefficient r is 0.953, which has a significance level greater than the 0.01. This shows that the control scores are predictive of the experimental error rates when the control scores are greater than 0.20. The correlation

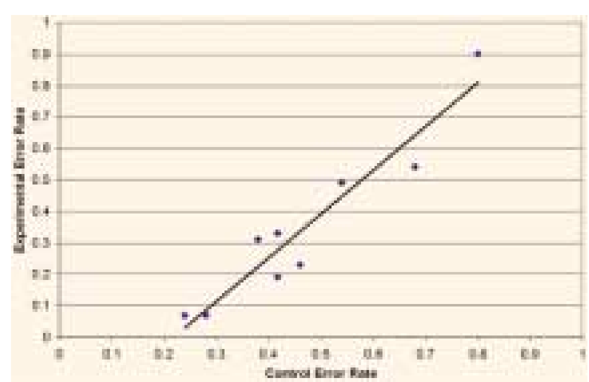

Figure 3. Best Curve Fit for Scatter Plot of Experimental vs. Control Error Rates for Experimental Scores with Controls Greater than 0.20.

Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR’02)
0-7695-1602-5/02 $17.00 © 2002 **IEEE**

coefficient for algorithms with control scores less than 0.20 was 0.283. Two possible explanations for the scores not being correlated are that performance is saturated or that for low error rates PCA-based algorithms are not appropriate control algorithms. The results of this analysis are summarized in table 2.

By comparing the relative performance of the control and experimental algorithms from the experiments in the meta-analysis paper suite and the FERET results, it can be seen that the majority of experiments in face recognition papers have concentrated on a relatively easy task. By reporting performance on data sets that saturate performance levels, it is hard to demonstrate significant breakthroughs in automatic face recognition, what is referred to as "diagnosticity" in behavioral decision research [37].

5. Conclusions

In this meta-analysis, we have identified areas of methodological problems in the manner face recognition experiments are performed and presented.

We have shown that there is strong correlation between the performances of control and experimental algorithms' scores, and therefore control scores are predictive of experimental algorithm performance. This strong correlation raises three questions for future investigations. First, why is improvement in performance of experimental algorithms only incremental over the control scores? Second, could one detect breakthroughs in face recognition through performance of an experimental algorithm that is not predicted by a control score? Third, because of the strong correlation, are all the algorithms using essentially the same information to perform recognition?

The majority of experiments in face recognition have concentrated on problems that have already been solved. These results have experimental error rates less than 0.10 and control algorithm error rates less than 0.20. As a result, experimental algorithm performance levels have been saturated, making robust comparisons between various algorithms impossible.

Based on the results of this analysis, we recommend that researchers concentrate on face recognition problems that are harder, as defined by the image sets in the experiments and the performance by a control algorithm. Researchers should use a standard control algorithm (such as PCA) on their test image set to determine its difficulty level and avoid easy tests for algorithms for which solutions already exist.

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 performance of a new face recognition algorithm. In the absence of an independent evaluation, the performance of control 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 control scores that fall above and below 0.20. The best control 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 control algorithms into an experimental method, it is necessary to establish accompanying standard evaluation protocols and image sets. This allows researchers to assess performance of new algorithms using established methodologies. This idea is similar to that pursued by the ORL sequence of experiments. What is suggested here would represent 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 major problem area is how authors describe the experiments and present the results. Only twelve out of the 24 papers in this study provided complete documentation. More attention needs to be paid to the details of the experimental design and careful reporting of results by authors, reviewers, and editors. Attention to these details will make for papers that

Table 2. Correlations Between Experimental and Control Error Rates.

	All Algorithms with a Control	Algorithms with Control Error Rates >0.20	Algorithms with Control Error Rates <0.20	Figure Number
Number of experimental algorithms	33	9	24	2
Correlation value r	0.932	0.953	0.283	3
Level of significance	<0.01	<0.01	Not significant	N/A

are more readable and will allow researchers to fully evaluate the contribution of a new algorithm (including performing meta-analysis) as well as independent replication of published performance results.

The face recognition and computer vision communities need to establish documentation standards for experiments. As a starting point for such a discussion, we recommend that each paper should report the following:

- Which database was used for the experiment as well as information about the database if it is one that the authors created;
  - performance percentages for at least the top rank (one) score in addition other scores the authors want to highlight the authors' experimental algorithm,
  - any variations of the experimental algorithm,
  - PCA or eigenface as a current control algorithm, and
  - any other control algorithm that the authors' chose to implement.
- The total number of images and the total number of subjects (or classes) for each of the following data sets:
  - the probe set,
  - training set, and
  - gallery set. If the gallery and training sets are not the same, state how they differ.
- What pose (frontal, profile, etc.) was tested in the experiment in the gallery versus probe image sets.

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*, 30, pp. 301-321, 2001.

[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.

[3] P.J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, Vol. 16, No. 5, pp. 295-306, 1998.

[4] M.S. Bartlett, H.M. Lades, and T.J. Sejnowski, "Independent component representations for face recognition," *Proceedings of the SPIE Symposium on Electronic Imaging: Science and Technology; Conference on Human Vision and Electronic Imaging III*, San Jose, California, January 1998.

[5] I.J. Cox, J. Ghosn, and P.N. Yianilos, "Feature-Based Face Recognition Using Mixture-Distance," *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 209-216, 1996.

[6] K. Etemad and R. Chellappa, "Discriminant Analysis for Recognition of Human Face Images," *Proceedings, International Conference on Audio- and Video-based Biometric Person Authentication*, pp. 127-142, 1999.

[7] G.G. Gordon, "Face Recognition from Frontal and Profile Views," *International Workshop on Automatic Face- and Gesture-Recognition*, Zurich, pp. 47-52, 1995.

[8] S. Lawrence, C.L. Giles, A.C. Tsoi, and A.D. Back, "Face Recognition: A Convolutional Neural-Network Approach," *IEEE Transactions on Neural Networks*, Vol. 8, No. 1, pp. 98-113, January 1997.

[9] S.Z. Li and J. Lu, "Face Recognition Using the Nearest Feature Line Method," *IEEE Transactions on Neural Networks*, Vol. 10, No. 2, pp. 439-443, March 1999.

[10] S. Lin, S. Kung, and L. Lin, "Face Recognition/ Detection by Probabilistic Decision-Based Neural Network," *IEEE Transactions on Neural Networks*, Vol. 8, No. 1, pp. 114-132, January 1997.

[11] C. Liu and H. Wechsler, "Comparative Assessment of Independent Component Analysis (ICA) for Face Recognition," *Proceedings, International Conference on Audio- and Video-based Biometric Person Authentication*, pp. 211-216, 1999.

[12] C. Liu and H. Wechsler, "Evolutionary Pursuit and Its Application to Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 6, pp. 570-582, June 2000.

[13] C. Liu and H. Wechsler, "Robust Coding Schemes for Indexing and Retrieval from Large Face Databases," *IEEE Transactions on Image Processing*, Vol. 9, No. 1, pp. 132-137, January 2000.

[14] S.M. Lucas, "Continuous n-tuple classifier and its application to face recognition," *Electronics Letters*, Vol. 33, No. 20, pp. 1676-1678, 25 September 1997.

[15] B. Moghaddam, C. Nastar, and A. Pentland, "Bayesian Face Recognition using Deformable Intensity Surfaces," *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 638-645, 1996.

[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,” *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.

[35] P.J. Phillips, D. Blackburn, and M. Bones, *FRVT 2000 Report. 2001*. Technical Report, http://www.dodcounterdrug.com/facialrecognition.

[36] J.L. Barron, D.J. Fleet, and S.S. Beauchemin, “Performance of Optical Flow Techniques,” *International Journal of Computer Vision*, 12:1, pp. 43-77, 1994.

[37] B. Fischhoff and R. Beyth-Marom. “Hypothesis evaluation from a Bayesian perspective.” *Psychological Review*, 90, pp. 239-260, 1983.