Transformation, Ranking, and Clustering for Face Recognition Algorithm Comparison

Stefan Leigh*, P. Jonathon Phillips,* Patrick Grother*,
Alan Heckert*, Andrew L. Rukhin†, Elaine Newton‡
Mariama Moody*, Kimball Kniskern*, Susan Heath*

* National Institute of Standards and Technology, Gaithersburg, MD 20899
†University of Maryland Baltimore County, Baltimore, MD 21250
‡Rand Corporation, Pittsburg, PA 15213

Abstract

The performance of face recognition algorithms is recently of increased interest. Empirical analyses of algorithms have traditionally been limited to rank-based scores such as cumulative match and receiver operating characteristics. This restriction to performance measures based on rank-based statistics arises because it is not possible to directly compare similarities output by algorithms. This paper presents the Phi-PIT transformation that makes it possible to compare such heterogeneous outputs, and allows a large body of classical statistical methods to be used to measure and analyze performance. These statistical techniques inclucde multiple comparison techniques, analysis of variance (ANOVA), and regression techniques. This paper presents ANOVA, graphical ANOVA, and Student-Newman-Keuls clustering analyses of the transformed outputs of fifteen face recognition algorithms.

1 Introduction

The last decade has seen significant advances in face recognition technology. However, these advances have not been accompanied by a clear understanding of why some recognition algorithms perform better than others. There is little consensus on which factors influence performance or on their relative importance. For example, how much of the observed variation in performance is attributable to the algorithms themselves? How much is due to the subjects or to the images? How much is due to subtle interactions among images and algorithms? What are the effects of training sets on performance? Credible, quantifiable answers to these questions are necessary for advances in face recognition.

Statistical techniques seem to offer a rich means for investigating performance properties of algorithms. The community acknowledges the utility of common image sets for comparing algorithmic performance, but the effort is complicated by the fact that algorithms estimate and rank identities using similarity scores. The scales that different algorithms employ for matching are distinct, and do not immedidately lend themselves to direct comparison.

Thus algorithm performance has, to date, been based on rank statistics; most frequently used are the cumulative match characteristics (CMC) curves for identification performance [14], and receiver operator characteristics (ROC) curves for verification performance.

Data generally fall into three classes: ratio, ordinal, nominal. Nominal data are named categories such as red, white, and blue, upon which arithmetic cannot be meaningfully performed. Ordinal data are typically represented as integers, with an implied or explicit relative ranking among the numbers: e.g., 10 is better than 9 is better than 8, etc. However, there is no information concerning how much better 10 is than 9, or 9 is than 8. Ratio data are real numbers and inherit the rich arithmetic and ordering structure of the real line.

One can always pass from ratio to ordinal scale by ranking the ratio scale numbers given by observation or experiment. Rank-based statistics have the mathematical virtue that, if the characteristics one seeks to compare *are invariant under monotonic transformation*, then "the best that one can do" is to compare via ranks [2]. But ranks have the very obvious potential defect of "throwing away information" in the passage from the richness of ratio scale similarity scores to the leaner ordinal scale of ranked scores.

If one could directly compare similarity scores among

*Please direct correspondence to P. Jonathon Phillips at jonathon@nist.gov.

algorithms *in their original ratio scale*, or transformed version thereof, it would open up the field to the use of a large existing corpus of statistical analysis tools for performance set ranking, clustering, and general comparisons. The list would include the many techniques of Multiple Comparisons [5], Analysis of Variance (ANOVA) [4] and variants (Multivariate ANOVA, General Linear Models, Mixed-Effects Models), regression techniques [1] ((Multi)linear, Nonlinear, Logistic, All Possible Subsets), and a host of multivariate techniques such as MultiDimensional Scaling, Discriminant Analysis, and Clustering; see, for example, [6].

Thus this paper discusses a transformation of similarities obtained from different algorithms tested on common sets of images that permits direct comparison *in a ratio scale*. We illustrate its use via graphical one-way ANOVA (boxplots, [9]) and a simple, easily interpretable graphical Multiple Comparison technique called Student-Newman-Keuls [5].

2 Image Selection

Face recognition algorithms estimate the identity of a person in a facial image "degree of match" between unknown subject *probe* and enrolled *gallery* image as "similarity scores". Performance of recognition algorithms can be obtained by post-hoc analysis of the matrix of similarity scores generated by comparing all probe images with all gallery images. Thus the similarity score, s_{ij}, is the result of comparing the i-th gallery and j-th probe images. This matrix of similarity scores and the identities of the subjects are the sole inputs necessary for computing recognition performance. For example, an ROC curve can be obtained by counting the relative numbers of matching and non-matching elements above a threshold t.

The similarity s_{ij} is a *match score* if the i-th gallery and j-th probe images are of the same person; a similarity is a *non-match* score if the images are of different people. A set of similarity scores is *homogeneous* if all elements are match scores. Likewise, a set is *heterogeneous* if all elements are non-matches.

The analysis performed in this paper is based on the Sep96 FERET evaluation [14], in which images from 1196 persons were used. We restrict attention here to those subjects for which the data contains at least eight images of the same human subject. The IDs of the FERET subjects meeting this criterion are: 93, 108, 182, 383, 468, 469, 547, 556, 588, 660, 705, 706, 708, 711, 717, 722, 744, 745, 751, 752 and 770. They are renumbered from 1 in the figures.

Each algorithm evaluated under the Sep96 FERET protocol generated 12,690,568 similarity scores corresponding to pairwise comparisons of images in sets of size 3,816 and 3,323.

Index	Name
1	baseline-cor
2	baseline-ef
3	ef-angle
4	ef-Mahalanobis-angle
5	ef-L1
6	ef-L2
7	ef-Mahalanobis
8	ef-Mahalanobis-L1
9	ef-Mahalanobis-L2
10	Excalibur
11	MIT-Mar-95
12	MIT-Sept-96
13	MSU
14	UMD-Mar-97
15	USC-Mar-97

Table 1. The fifteen algorithms and their indices.

The outputs of fifteen algorithms are considered. Six were independently developed and evaluated in the Sep96 FERET evaluations. They are: MIT-Mar95 [10], MIT-Sep96 [11], MSU [17], UMD-Mar97 [16, 17], USC-Mar97 [7], and one developed by Excalibur Corp. Two more, baseline-ef and baseline-cor, were baseline algorithms from the Sep96 FERET evaluations [14]. The final seven algorithms are due to Moon and Phillips [12], are prefixed by ”ef-”, and are implementations of the principal components analysis (PCA) face recognition algorithm. They differ only in the distance metric used in the nearest neighbor classification. In all the figures that follow, the algorithms are indexed by the numbers given in table 1.

3 The PIT and Phi-PIT Transformations

It is often the case, with real world data, that in order to make meaningful inferences on the structure of the data, one must subject elements of the data to transformation [3]. Richter scale for earthquakes, the pH scale for acidity, decibels for intensity of sounds are all common examples of physical data numbers re-expressed by - logarithmic - transformation in order to facilitate understanding and operation.

In our situation each algorithm yields similarity scores on different scales as a result of using different normalizations and distance metrics. This is shown in Figure 1, which plots the empirical cumulative distribution function (ECDF) of the MSU, UMD97, and USC97 algorithms. The ECDFs were generated from all 12,690,568 similarity elements. For these three algorithms, the range of the similarities scores was in $[0, 1]$. Figure 2 presents a boxplot of the similarity scores for all fifteen algorithms. Figures 1 and 2

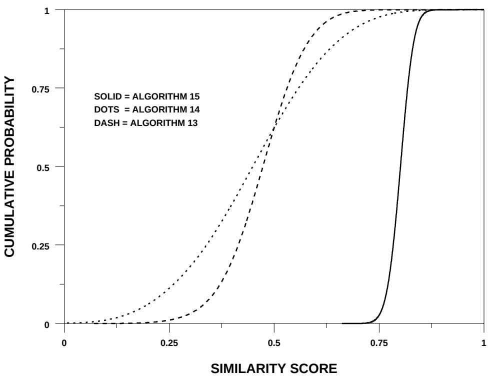

Figure 1. Empirical CDFs, for the MSU, UMD97, and USC97 algorithms.

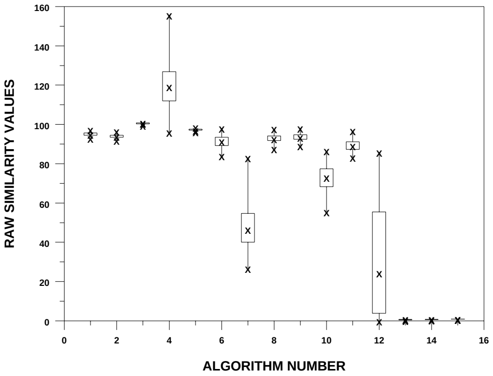

Figure 2. Boxplot of raw similarity scores.

show that the arbitrary ranges and distributions of the raw similarity scores do not allow for direct comparison.

Meaningful comparison of algorithms is only possible if the similarity scores are put on a single scale. A coherent renormalization of similarity values can be achieved by transforming them by their ECDF. The application of the ECDF constitutes the Probability Integral Transform (PIT) in which numbers distributed according to a parent CDF are transformed into uniformly distributed numbers on the unit interval. So if the ECDF of the k-th algorithm is $\hat{F}_k$, then similarity scores coming from algorithm k are transformed according to:

$$u_{ij} = \hat{F}_k(s_{ij}). \tag{1}$$

The result is similarity values that are uniformly distributed on the unit interval. Because all values are transformed by their native ECDF onto the unit interval the scores are now directly comparable. The ECDFs are estimated from the 12,690,568 similarities scores.

A second recommended step in the transforming process is to apply the inverse cumulative function of the standard Gaussian (normal), Φ^{-1}, to the PIT-scores obtained in step one:

$$p_{ij} = \Phi^{-1}(\hat{F}_k(s_{ij})) = \Phi^{-1}(u_{ij}) \tag{2}$$

This transformation takes the uniformly distributed PIT-scores and converts them to normally distributed Phi-scores. A Gaussian form for score expression is selected because many classical statistical comparison techniques assume approximate normality of datasets being compared.

Transformation to other distributional forms is easily achieved. Many standard non-uniform random number generators are based on appropriate transformation of uniformly distributed numbers. To transform and compare to lognormality, for example, the transformation

$$e_{ij} = \exp(\Phi^{-1}(\hat{F}_k(s_{ij}))) = \exp(p_{ij}) \tag{3}$$

should be employed.

Each of the above transformations is monotone, so ranks are preserved. Thus general rank order statistics, such as CMC and ROC scores are invariant.

4 Score Comparison via Boxplots

Having coherently transformed all scores from an algorithmic comparison experiment into a Gaussian framework, retaining the original *ratio* scale of the data, we are now at liberty to rank, cluster, and generally compare sets of algorithmic scores on comparable Gallery-Probe match problems using the whole panoply of classical statistical methods.

A boxplot is a graphic technique for visually characterizing the distribution of a data set, or sets of data. It is an easily implemented schematic, used to compare the empirical distributions represented by batches of numbers. It can be thought of as constituting a visual one-way ANOVA. Location, spread, and extreme information for each batch are embedded in the graphical display. This allows meaningful comparison of distributional information through rapid assessment of the alignment or misalignment, ranking, and clustering of median values and boxes, and differences in spread.

Here, the batches of numbers being compared are raw or suitably transformed scores, indexed by algorithm or image subject.

Important features of the boxplot are: (1) the width of each box is proportional to the data set size; (2) the median value of the data, used as an indicator of location because of its resistance to outliers, is identified by the X; (3) the interquartile range (the range from the 25^{th} to 75^{th} percentile, i.e. the “middle half”) of the data is represented by the body of the box; (4) the extremes (minimum and maximum) are represented by the whiskers (ends of the straight lines) projecting out of the box.

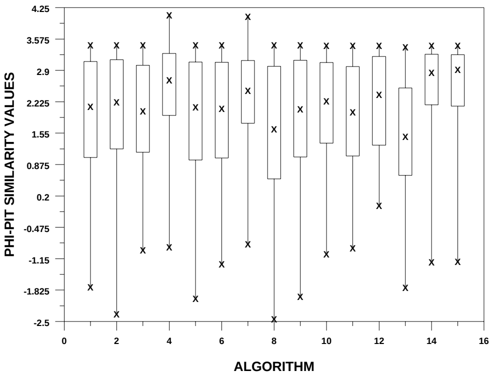

Figure 3. Boxplot of Phi-PIT similarity scores by algorithm.

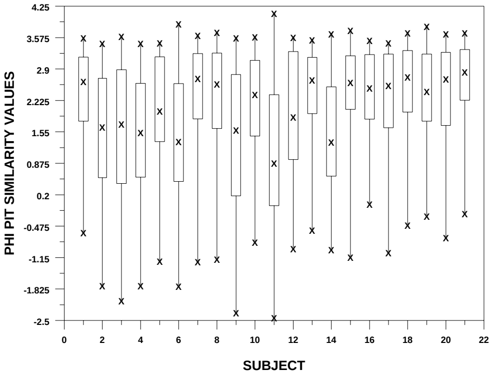

Figure 4. Boxplot of Phi-PIT scores by subject showing more pronounced subject effect in terms of median slippage than original untransformed data.

As described in section 2 there are at least eight images for each subject. In general let each subject have L_i images, and let P_i^k be the set of homogeneous scores for person i reported by algorithm k. If the similarity scores for algorithm k are symmetric, there are $L_i(L_i - 1)/2$ different similarity scores. If the scores are asymmetric, each pair are averaged to produce $L_i(L_i - 1)/2$ scores. Let P^k be the union of all 21 sets of homogeneous scores for algorithm k; i.e., $P^k = \cup_i P_i^k$. Figure 3 is a boxplot of the P^k's for all fifteen algorithms. Because the scores have been transformed to be on the same scale it is meaningful to compare them across algorithms. The majority of the scores are above zero because the lower non-match scores have been omitted. In fact, a correlation between the relative height of an algorithms boxplot with the other algorithms and its performance on traditional measures such the CMC and ROC is expected.

Figure 3 shows algorithm effects: discernible variation in homogeneous similarity scores across the algorithms. But what is the difference in distribution of homogeneous scores among the 21 subjects. This plot assesses whether some people harder to recognize than others. In the psychological literature this is connected to the typical vs. atypical people issue, and it has been shown that humans have a harder time recognizing typical people [8, 13, 15]. Do some subjects display a larger range of similarity scores?

The first step is to collect all the similarity scores for each person. Let S_i be all the similarity scores for all fifteen algorithms for person i; i.e., $S_i = \cup_k P_i^k$. Figure 4 contains the boxplots for all 21 S_i's. Figure 4 shows a large variation in the distribution of similarity scores among the 21 subjects, in terms of median, and interquartile levels, and range of each distribution.

Figures 3 and 4 show algorithm and subject effects in isolation. Now we look at interaction between algorithm and subject effects. In Figure 4, each subject has a range of scores, however, within a subject we do not have the distribution broken out by algorithm. If for subject 1, algorithm j has the highest ranked median, is the same true for the remaining algorithms? One can look at how the ranking of means for P_i^k vary using graphical analysis of variance (GANOVA). Means are used because it is not possible to simultaneously display the entire set of homogeneous similarities in a way that overall structure can be appreciated.

Figures 6 and 5 show the GANOVA of the interaction between the algorithms and subject effects, which we now explain. The GANOVA plots Phi-PIT scores against an index denoting algorithm or subject while each trace corresponds to one subject or algorithm. Let $\overline{P_i^k}$ denote the mean of the P_i^k's. Trace k, corresponds to algorithm k and connects the points $\{\overline{P_1^k}, \overline{P_2^k}, \ldots, \overline{P_{21}^k}\}$. This shows how the means of the P_i^k's vary by subject and algorithm. Stratification of the traces indicates a clear factor effect. Any criss-crossing of the traces corresponds to interaction between algorithm and subject manifested by variation in levels of one factor across different levels of the other. Figure 6 shows some stratification, while also showing some criss-crossing. The amount of stratification and criss-crossing is dependent on the subjects. Figure 5 shows less stratification indicative of less pronounced subject effect.

5 Analysis of Phi-PIT-scores

The graphical techniques of the previous section that do not make any distributional assumptions. However, without distributional assumptions statistical inference cannot be performed. Because the Phi-PIT transformation yields a data set of 12,690,568 values that are normally distributed, it is possible to perform statistical inference tests that require the data with a Gaussian distribution.

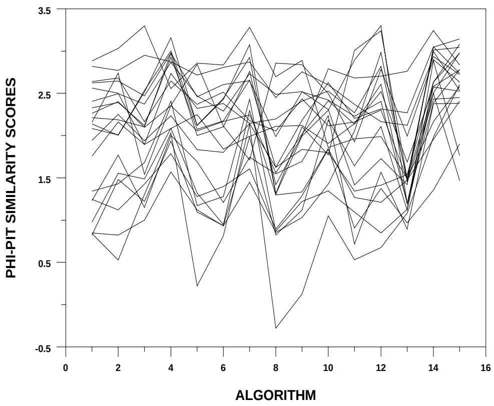

Figure 5. Graphical ANOVA of Phi-PIT similarity values. Each trace corresponds to one subject algorithm.

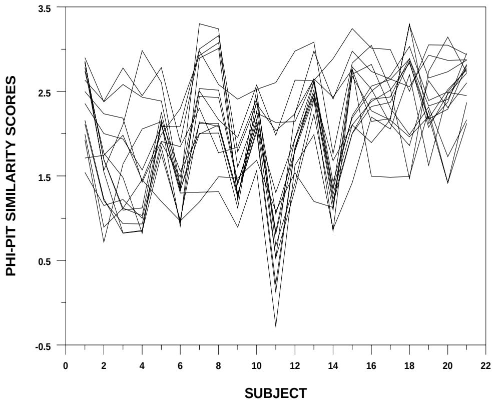

Figure 6. Graphical ANOVA of Phi-PIT similarity values. Each trace corresponds to one subject. The plot shows both the strong stratification indicative of algorithm effect, and criss-crossing suggestive of modest interaction between algorithm and subject.

First ANOVA can be used to formally analyze the interaction between algorithm and subject effects that were graphically studied above. The results of the ANOVA appear in table 2. The F-values and near-zero P-values indicate unambiguous algorithm and subject effects, and less significant algorithm-subject interaction.

The standard global F-test for equality of means associated with the oneway ANOVA is a one-shot test for assessing the "homogeneity" (equality) of the means of groups being compared. If the test fails, as it has here, it becomes of natural interest to consider what subgroupings of means might exhibit homogeneity. There exists a large corpus of such techniques in the statistical literature, which goes under the name of Multiple Comparisons.

Effect	SSQ	DF	MSQ	F-value	P-value
Algorithm	1671.0	14	119.4	126.2	0.000
Subject	3645.9	20	182.3	192.8	0.000
Interaction	2533.9	280	9.1	9.6	0.000

Table 2. ANOVA table for replicated 2-factor fixed-effect with interaction model for Phi-PIT scores. The small P-values indicate significant algorithm and subject effects and algorithm:subject interaction.

Student-Newman-Keuls (SNK) is a standard Multiple Comparisons procedure designed in part to protect against the risk of inflated Type I error associated with procedures of the "all possible t-tests" type. (Type I error means erroneous rejection of null hypotheses of equality of clusters of means.) If the mean Phi-PIT scores are computed for each of the algorithms being considered, a natural (t-) statistic for pairwise comparison of means is the (absolute value of) the difference between the means denominated by a standard error of that difference. SNK concentrates on the largest such statistic, the so-called Studentized range, from all such possible pairwise comparisons. SNK is a stepdown procedure for testing all possible subset homogeneity (equality of means) hypotheses based on the distribution of the largest such. The procedure starts by checking whether all k algorithm means can be clustered; then whether any of the (k-1)-fold sets of means can be clustered; then any of the (k-2)-fold sets etc. At each level the homogeneity (equal means) hypothesis is tested against the appropriate Studentized range. The resulting SNK schematic in figure 7 is an easily understood ranking and clustering of the groups (here: algorithms) being compared. The numbers represent the mean Phi-PIT scores for the algorithms indicated. The figure shows that an SNK cluster of the two best algorithms in the top rank, a clear clustering of the worst algorithms at the bottom, and groupings of "in-between" algorithms in the middle.

6 Discussion and Conclusion

The paper employs the Phi-PIT transformation of the heterogeneous outputs of face recognition systems. It permits direct comparison of similarity scores across systems in the ratio scale. This normalization immediately enables the use of a large number of standard statistical procedures for comparison of algorithms. Further the gaussianity of the transformed scores allows techniques that require it to be invoked, such as ANOVA or SNK. The restriction to monotonic transformations preserves the rank-order statistics of the data. This has the desirable property of leaving traditional scoring metrics unchanged.

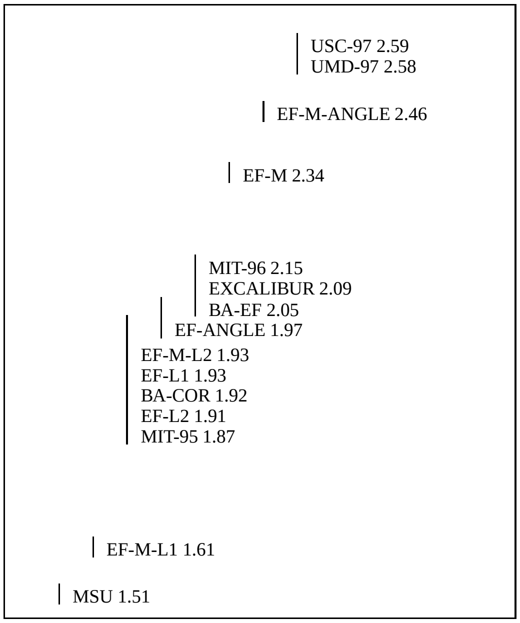

Figure 7. SNK ranking and clustering of algorithms. The vertical displacement is linear with the mean Phi-PIT value; the horizontal displacement and vertical lines give the clustering. Extended lines indicate cluster overlap.

References

[1] N. R. Draper and H. Smith. *Applied Regression Analysis*. John Wiley & Sons, New York, 3rd edition, 1998.

[2] J. Hajek and Z. Sidak. *Theory of Rank Tests*. Academic Press, New York, 1967.

[3] D. Hoaglin. Transformations in everyday experience. *Chance*, 1(4):40–45, 1988.

[4] D. Hoaglin, F. Mosteller, and J. Tukey. *Fundamentals of Exploratory Analysis of Variance*. John Wiley, New York, 1991.

[5] Y. Hochberg and A. C. Tamhane. *Multiple Comparison Procedures*. John Wiley & Sons, New York, 1987.

[6] R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, Englewood Cliffs, N.J., 1992.

[7] M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. on Computers*, 42:300–311, 1993.

[8] L. Light, F. Kayra-Stuart, and S. Hollander. Recognition memory for typical and usual faces. *Journal of Experimental Psychology: Human Learning and Memory*, 5:212–228, 1979.

[9] D. R. McNeil. *Interactive Data Analysis: A Practical Primer*. John Wiley & Sons, New York, 1977.

[10] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. *IEEE Trans. PAMI*, 19(7):696–710, 1997.

[11] B. Moghaddam and A. Pentland. Beyond linear eigenspaces: Bayesian matching for face recognition. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*, pages 230–243. Springer-Verlag, Berlin, 1998.

[12] H. Moon and P. J. Phillips. Computational and performance aspects of PCA-based face-recognition algorithms. *Perception*, 30:303–321, 2001.

[13] A. J. O'Toole, K. A. Deffenbacher, D. Valentin, and H. Abdi. Structural aspects of face recognition and the other-race effect. *Memory & Cognition*, 22(2):208–224, 1994.

[14] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:1090–1104, 2000.

[15] T. Valentine and V. Bruce. The effects of distinctiveness in recognising and classifying faces. *Perception*, 15:525–536, 1986.

[16] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. pages 336–341, 1998.

[17] W. Zhao, A. Krishnaswamy, R. Chellappa, D. Swets, and J. Weng. Discriminant analysis of principal components for face recognition. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*, pages 73–85. Springer-Verlag, Berlin, 1998.