UNIVERSITE DE
GENEVE

CENTRE
UNIVERSITAIRE
D'INFORMATIQUE
GROUPE VISION

Date: January 30, 1998
N^{o} 98.01

TECHNICAL REPORT

VISION

Generalization performance of factor analysis techniques used for image database organization

David McG. Squire

Computer Vision Group
Computing Science Center, University of Geneva
24 rue du Général Dufour, CH - 1211 Geneva 4 SWITZERLAND
e-mail: squire@cui.unige.ch

Abstract

The goal of this paper is to evaluate the generalization performance of a variety of factor analysis techniques in an image database environment. Factor analysis techniques, such as Principal Components Analysis, have been proposed as means of reducing the dimensionality of the data stored in image retrieval systems. These techniques compute a transformation which is applied to vectors of image features to produce vectors of lower dimensionality which still characterize the original data well. Computing such transformations for very large numbers of images is computationally expensive, especially if this calculation must be repeated each time new images are added to the database. It is to be hoped, therefore, that a transformation computed using a subset of all possible images will perform well when applied to images not used in its derivation. To evaluate this generalization ability, we measure the agreement between partitionings of image sets computed using such transformations with those produced by human subjects.

Keywords: image databases, clustering, factor analysis, generalization, agreement

1 Introduction

Millions of people now use multimedia documents daily: on the world wide web, in the electronic preparation of documents, and in the everyday use of computers. Multimedia documents are distinguished by the presence of images. There is thus a great need for systems that allow users to create, manage and query image databases in an efficient and accurate manner. The attachment of textual labels to images is considered inadequate for these purposes. There is thus great interest in content-based image retrieval systems (CBIRSs). A CBIRS retrieves images from a database based on their similarity to a query image or sketch [1, 2].

One of the problems faced by CBIRS designers is the vast quantity of data that is available to them. A great variety of image features have been proposed for use in CBIRSs, including frequently-used colour features such as colour histograms [1, 3, 4], texture features [5, 6], shape features [7, 8], and a variety of region and edge statistics [3, 9]. In short, a large number of features is available, and it is difficult to know their dependencies *a priori*.

Factor analysis provides a means of addressing this problem. Factor analysis has been applied to text retrieval, where it is usually known as "Latent Semantic Indexing" [10]. It has also been used for image retrieval by several authors, *e.g.* [4].

In this paper, we investigate the generalization capacity of factor analysis techniques in an image database system: how well do image similarity measures derived using factor analysis generalize to images not included in the factor calculation?

2 Factor analysis techniques

Factor analysis techniques provide a means of representing large sets of numerical data in spaces of lower dimensionality, by finding a new, ordered, set of orthogonal axes, so that the sum of the norms of the projections of the data onto these axes is maximized, according to some metric.

Perhaps the best-known such method is Principal Component Analysis (PCA), which uses the covariance of the mean-centered data as the metric, thus providing factors that maximally preserve the *variance* of the original data. If the raw data for each feature is first normalized by its standard deviation (NPCA), then the *correlation* between data points becomes the metric for the analysis.

In Correspondence Analysis (CA) [11], the coordinates of the data points are defined so that the Euclidean metric in feature space corresponds to the χ^2 distance between the points, and the analysis is thus in terms of the *independence* of the data.

All these techniques lead to a set of linear equations which map the (perhaps already transformed) original, high-dimensional feature vectors to a factor space. The feature vectors are projected onto the eigenvectors of a matrix. The corresponding ordered eigenvalues indicate the degree to which each eigenvector contributes to the approximation of the original space spanned by the data. Since there are usually dependencies between the original features, it is usually possible to obtain a high quality approximation of the data space retaining only the first few eigenvectors. The techniques can thus be used to find mappings from high-dimensional feature spaces to lower dimensional factor spaces, such that some statistic of the original data is still well approximated.

3 Agreement between partitionings of an image set

In order to evaluate the performance of image database organization systems, we have developed a measure of the agreement between two partitionings of a set of images into unlabeled subsets, based on pair-wise subset membership comparisons. This differs from the precision and recall performance measures used for evaluating a system's responses to queries.

We are interested in such partitionings since they provide a means of restricting the number of images which must be searched during a query. Also, by evaluating the way systems and humans partition a set of images into subsets, we effectively compare their similarity judgments for all images in the set simultaneously. In earlier studies [9, 12], it emerged that random partitionings can have significant chance agreement. We have shown how a better, chance-corrected, agreement measure can be defined. We call this agreement measure κ_B; the mathematical details may be found in [12] (the main interest is the calculation of expected chance agreement). Based on Cohen's kappa statistic [13], it is 0 for no improvement over change agreement, and 1 for perfect agreement. It is defined as:

$$\kappa_B = \frac{\text{observed agreement} - \text{expected chance agreement}}{1 - \text{expected chance agreement}}. \tag{1}$$

4 Experiments

We based our study on a set of 500 unconstrained colour provided by Télévision Suisse Romande. From these images, two subsets of 100 images were selected randomly. For each image, 15 features were extracted, based on colour statistics, segment and arc lengths, arc radii and region intensity variances [14]. Three factor analysis techniques (CA, NPCA and PCA) were applied to these features, for each set of 100 images and for the full set of 500. The eigensystems describing the resultant mappings were saved. The images were then clustered into a binary tree, using Ascendent Hierarchical Clustering [15], for each factor analysis type, retaining either 2 or 4 factors. For each of the two subsets of 100 images, the images were also clustered after their features were transformed using the eigensystem derived from the other subset, as well as that derived from the full 500 images.

Human subjects were also asked to partition the sets of 100 images into at most 8 subsets. This was done using a computer program which initially presented each subject with the source image set, and 8 empty sets. Images could be dragged from any image set and dropped in another image set using the mouse. When all images from the source image set had been assigned to subsets, the partitioning could be saved.

Subjects were given a brief demonstration of how the program worked, and told that the notion of image similarity was entirely their choice. The task was performed by 9 members of the computer vision research group at the Université de Genève, who may be considered to be have some expert knowledge.

The third level of a binary tree contains (at most) eight classes. The machine clusterings were thus extracted from the binary trees at that level to facilitate their comparison with the human partitionings. The agreement between all the machine and human partitionings was then calculated for each image set.

5 Results

In order to put these results in context, one much first remark that the agreement between human subjects varied greatly. A summary appears in Table 1. Though very much greater than those expected by chance, these results show that it is unreasonable to expect any CBIRS to perform perfectly across all users.

	mean	median	std. dev.	κ_B min.	κ_B max.
Image set 1	0.3773	0.3625	0.0822	0.2225	0.5868
Image set 2	0.4383	0.4668	0.1082	0.2304	0.6557

Table 1: Statistics summarizing the agreement between human subjects, using κ_B, for image sets 1 and 2.

For each image set, we present, in Tables 2 and 3, the mean agreement between the human and machine partitionings for each of the factor analysis techniques, under the various conditions described above. The abbreviation "eigs. 1" indicates the eigensystem calculated using image set 1; "2 facts." indicates that two factors were retained for clustering, *etc.*.

It must be remembered that these are average values, computed using the agreements with 9 human subjects. As is clear from Table 1, the agreement between these humans was far from 100%. Also, the standard deviations of these figures were in the range of 0.01 to 0.05. The differences between the performances of the various techniques are often in this range.

	eigs. 1 2 facts.	eigs. 1 4 facts.	eigs. 2 2 facts.	eigs. 2 4 facts.	eigs. all 2 facts.	eigs. all 4 facts.
CA	0.1493	0.1218	0.1212	0.1170	0.1457	0.1218
NPCA	0.1368	0.1367	0.0864	0.0973	0.1188	0.1387
PCA	0.1350	0.1162	0.1421	0.1180	0.1157	0.1173

Table 2: Mean agreement between human and machine partitionings for image set 1, for various factor analysis types, eigensystems, and numbers of factors retained.

	eigs. 2 2 facts.	eigs. 2 4 facts.	eigs. 1 2 facts.	eigs. 1 4 facts.	eigs. all 2 facts.	eigs. all 4 facts.
CA	0.1146	0.1901	0.1311	0.1581	0.1328	0.1472
NPCA	0.1938	0.1468	0.1679	0.1796	0.1542	0.1508
PCA	0.1284	0.1320	0.1376	0.1450	0.1284	0.1321

Table 3: Mean agreement between human and machine partitionings for image set 2, for various factor analysis types, eigensystems, and numbers of factors retained.

For image set 1, the use of the eigensystems computed using image set 2 reduced performance in all cases except for PCA, and to values below those obtained using the eigensystems computed using all 500 images. For PCA the values were still within one standard deviation of the prior values. Using the eigensystems from all 500 images, there was almost no change compared to the values computed using the eigensystems from image set 1, and the small reductions in performance were all well within one standard deviation of the means.

For image set 2, the results are more mixed, though in 4 out of 6 cases, agreement was better using the eigensystems from image set 1 rather than that from image set 2 itself. Using the eigensystems computed from all 500 images, agreement was better or effectively unchanged in 4 out of 6 cases, with the differences again being small.

6 Conclusion

It is difficult to draw any firm conclusions from these experiments. It is clear that, although much less than 100%, the agreement between human partitionings of image sets is much greater than that between humans and machine partitionings. The average agreement between humans of 40.78% indicates that it is unrealistic to expect any CBIRS to satisfy all users. To do better, a CBIRS would have to learn the characteristics of particular users.

The gap between the average human/machine agreement over all permutations of techniques of 13.63% and the human/human average no doubt indicates that the features used as the raw data for the factor analysis techniques failed to capture important information used in human similarity judgments. Factor analysis techniques cannot extract information which isn't there. It is worth noting, however, that the clusters produced by these methods were pronounced "good" by several human observers – humans are all too good at finding reasons for images to be associated.

Using a subset of 20% of the total image set did not adversely affect agreement with human subjects. Image set 1 in our experiments appears to have been a slightly better subset than image set 2. Since the computation time for these techniques grows at least as the square of the number of images, the ability to use such a subset is is a significant advantage. A more exhaustive evaluation of generalization performance is desirable, but, as in most aspects of CBIRSs, the great difficulty obtaining ground truth in the form of human similarity judgments.

References

[1] A. Gupta and R. Jain, "Visual information retrieval," *Communications of the ACM*, vol. 40, May 1997.

[2] J. Ze Wang, G. Wiederhold, O. Firschein, and S. Xin Wei, "Wavelet-based image indexing techniques with partial sketch retrieval capability," in *Proceedings of the Fourth Forum on Research and Technology Advances in Digital Libraries*, (Washington D.C.), May 1997.

[3] A. K. Jain and A. Vailaya, "Image retrieval using color and shape," *Pattern Recognition*, vol. 29, pp. 1233–1244, August 1996.

[4] S. Sclaroff, L. Taycher, and M. La Cascia, "ImageRover: a content-based browser for the world wide web," in *IEEE Workshop on Content-Based Access of Image and Video Libraries*, (San Juan, Puerto Rico), June 1997.

[5] T. P. Minka and R. W. Picard, "Interactive learning using a "society of models"," in *Proceedings of the 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96)*, (San Francisco, California), pp. 447–452, IEEE Computer Society Press, June 1996.

[6] S. C. Orphanoudakis, C. E. Chronaki, , and D. Vamvaka, "I^2*net*: Content-based similarity search in geographically distributed repositories of medical images," *Computerized Medical Imaging and Graphics*, vol. 20, no. 4, pp. 193–207, 1996.

[7] S. Sclaroff, "Deformable prototypes for encoding shape categories in image databases," *Pattern Recognition*, vol. 30, April 1997.

[8] A. D. Bimbo and P. Pala, "Shape indexing by multi-scale representation," in *Image Databases and Multi-Media Search* (A. W. M. Smeulders and R. Jain, eds.), (Kruislaan 403, 1098 SJ Amsterdam, The Netherlands), pp. 43–50, Intelligent Sensory Information Systems, Faculty of Mathematics, Computer Science, Physics and Astronomy, Amsterdam University Press, August 1996.

[9] D. M. Squire and T. Pun, "A comparison of human and machine assessments of image similarity for the organization of image databases," in *The 10th Scandinavian Conference on Image Analysis* (M. Frydrych, J. Parkkinen, and A. Visa, eds.), (Lappeenranta, Finland), pp. 51–58, Pattern Recognition Society of Finland, June 1997.

[10] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *Journal of the Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[11] J. P. Benzécri, *L'analyse des Données; Tome II: l'Analyse des Corrrespondances.* Paris: Dunod, 1973.

[12] D. M. Squire and T. Pun, "Assessing agreement between human and machine clusterings of image databases," *Pattern Recognition*, 1998. (accepted for publication, to appear 1998).

[13] G. Dunn, *Design and analysis of reliability studies; the statistical evaluation of measurement errors.* 200 Madison Avenue, New York, NY 10016: Oxford University Press, 1989.

[14] T. Pun and D. M. Squire, "Statistical structuring of pictorial databases for content-based image retrieval systems," *Pattern Recognition Letters*, vol. 17, pp. 1299–1310, 1996.

[15] E. Diday, G. Govaert, Y. Lechevallier, and J. Sidi, "Clustering in pattern recognition," in *Digital Image Processing* (J.-C. Simon and R. Haralick, eds.), pp. 19–58, D. Reidel Publishers, 1981.