UNIVERSITE DE GENEVE

CENTRE UNIVERSITAIRE
D'INFORMATIQUE
GROUPE VISION

Date: November 24, 1997
N° 97.06

TECHNICAL REPORT

VISION

Using human partitionings of an image set to learn a similarity-based distance measure

David McG. Squire Thierry Pun[1]

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 pun@cui.unige.ch

[1] This work is supported by the Swiss National Foundation for Scientific Research (grant no. 2100-045581.95).

Abstract

In this paper our goal is to employ human judgments of image similarity to improve the organization of an image database for content-based retrieval. We first derive a statistic, κ_B for measuring the agreement between two partitionings of an image set into unlabeled subsets. This measure can be used both to measure the degree of agreement between pairs of human subjects, and also between human and machine partitionings of an image set. This provides a rigorous means of selecting between competing image database organization systems, and assessing how close the performance of such systems is to that which might be expected from a database organization done by hand.

We then use the results of experiments in which human subjects are asked to partition a set of images into unlabeled subsets to define a similarity measure for pairs of images based on the frequency with which they were judged to be similar. We show that, when this measure is used to partition an image set using a clustering technique, the resultant clustering agrees better with those produced by human subjects than any of the feature space-based techniques investigated. Finally, we investigate the use of machine learning techniques to discover a mapping from a numerical feature space to this perceptual similarity space. Such a mapping would allow the ground truth knowledge abstracted from the human judgments to be generalized to unseen images.

1 Introduction

The explosive growth of the world wide web means that millions of people now access multimedia documents daily. The use of digital images is also now standard practice in the preparation of paper documents. The distinguishing characteristic of multimedia documents is the presence of images, whether static or as components of a video sequence. 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 inadequate for these purpose, since identical images can be described in completely different ways, and controlled vocabulary indexing is now considered insufficient even in text retrieval systems. Consequently, there is significant 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]. There are now several commercial CBIRSs available, such as IBM's QBIC system [3], Excalibur Technology's Visual Retrieval Ware [4], Virage's Visual Information Retrieval engine [1], as well as systems for searching for images on the world wide web, such as ImageRover [5]. The emergence of commercial systems does not indicate that the technology is mature, only that the demand for it is very strong.

Current systems face great difficulties, due to the fact that *perceived image similarity* is both subjective and task-dependent. Image database organization, feature selection, efficient search and user-modeling remain difficult problems. In this context, we seek to enhance human usability of image archival and retrieval systems by integrating new methods of *image database organization*, and by using machine learning to incorporate *human similarity judgments* in this process. The resultant system should have a better measure of image similarity than those based solely upon image features.

In order to investigate human perception of image similarity, we have performed experiments to measure the agreement between a number of human partitionings of a set of images, as well as agreement between machine partitionings and human partitionings. 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. It emerges that random partitionings can have significant chance agreement. We have investigated the statistics of this agreement measure, and shown how a better, *chance-corrected*, agreement measure can be defined. The expected chance agreement can be large, especially for the small image sets often used to test such systems. It is *vital* to take this into account.

We found that agreement between humans is significantly better than that expected by chance, but much less than might have been anticipated. Agreement between human and machine partitionings is not as great. These results indicate that no single machine clustering technique can be expected to satisfy all database users. Nevertheless, our agreement measure can be used to aid the selection of image features and techniques for dimensionality reduction and clustering.

We then show how a collection of human partitionings of an image set can be used to define a ground truth similarity measure for each pair of images in this set. The use of this measure to create a machine partitioning gives better agreement with human partitionings than any other method tried, and indeed agrees with humans better, on average, than they do amongst themselves. Finally, we report a preliminary investigation into the use of machine learning to find a mapping between numerical image features and this similarity function.

2 State of the art

2.1 Features

It is acknowledged that semantic retrieval remains impossible, *i.e.* no existing system can retrieve all images of cats, regardless of colour, background and pose, from a large heterogeneous database. This difficulty can be partially avoided by using only images from restricted domains, such as industrial trademarks [6, 7] or marine animals [8]. In tackling the general problem, low-level image features are usually used, and an attempt is made to capture similarity using some function of these.

The most frequently used feature is colour [5, 7]. Similarity is defined as some distance between image colour distributions, the most common being the colour histogram [1, 9]. Many systems use texture features [10, 11], such as hierarchies of Gabor filters [12]; the Wold features [13] used in Photobook [14]; the coarseness, contrast, and directionality features used in QBIC [3]; others including covariance, correlation and entropy models [11] and wavelet-based decompositions [15]. The limitation of such global features is that images may have similar colour or texture statistics, but no semantic similarity, due to differing spatial distributions of these features.

Shape features are also often global (an image is assumed to contain one shape), and are thus most easily applied to restricted domains. Modal matching, for example, has been applied to fish, rabbits and machine tools [16]. Other shape-based approaches include multi-scale representation of curves [17], histograms of edge directions [7, 18], maxima of zero-crossing contours of curvature scale space images [8], and matching templates of shape components [19].

Global descriptors can be augmented by seeking features which retain spatial information, such as Daubechies' or Haar wavelet decompositions [2, 20]. Another approach is to segment the image into regions, from which features are extracted, such as color, size, location and relationships to other regions. This approach turns the image retrieval problem into labeled graph matching, which is known to be NP complete.

2.2 Similarity

CBIRSs aim to return images which, according to human perception, are *similar* to a query image. Remarkably, few of these systems consider what similarity means in the context of human usage. Those that do report that human judgments of similarity noticeably differ (*e.g.* [8]). Typically, images are represented as points in a multidimensional feature space. A metric defined on this space is used to measure dissimilarity between images: images close to the query, according to this metric, are *similar* to the query. That human and machine notions of similarity may be very different is rarely discussed.

It is often implied that given the "right" features (an appropriate colour space [5, 9], texture features "corresponding to human perception" [13]), proximity in feature space *must* correspond to perceptual similarity. There are several reasons to doubt this, the most fundamental of which is the *metric assumption*. There is psychophysical evidence that human similarity judgments do not obey the requirements of a metric: self-identity, symmetry and the triangle inequality [21].

Some authors have addressed the fact that distance in feature space is not equivalent to perceptual similarity. For example, self-organizing maps have been used to cluster texture features according to class labels provided by human judgments of texture similarity [12]. Minka and Picard report a system which learns groupings of similar images from positive and negative examples provided by users during query sessions [10].

3 Image similarity and agreement between partitionings of a set

It is difficult to make an objective assessment of the performance of CBIRSs because image retrieval researchers lack large sets of images for which the similarity "ground truth" is known. In contrast, text-based document retrieval researchers frequently use data from the same large, expert-classified datasets, and quantitative comparisons between document retrieval systems are made, notably in the TREC conference series.[1]

In order to investigate human judgments of image similarity, we asked human subjects to partition a set of colour images with unconstrained content into a number of subsets, with no prompting or guidance. A method for assessing the agreement between partitionings produced by pairs of subjects was developed [22, 23], based on statistical measures of reliability well-known in medical and psychological research [24, 25].

We also used a variety of machine systems to cluster the same set of images that was presented to the human subjects. We then computed the agreement between the machine and the human partitionings. Averaged over all humans, this provides a measure of the degree of overlap of each machine measure of image similarity with the common human measure. This measure can be used to rank that machine systems, and thus to chose between them. The average agreement between pairs of humans gives an indication of the best performance that could be expected of *any* machine partitioning.

3.1 The κ_B statistic

To measure the agreement between two partitionings of an image set, we consider pairs of images individually (since the subsets are unlabeled). Consider the set of images $\Phi = \{I_1, \ldots, I_N\}$. Two subjects, A and B, independently partition Φ into M subsets. We

[1] **T**ext **RE**trieval **C**onference – Further information and proceedings are available at: "`http://trec.nist.gov/`".

will call the resultant sets of subsets of Φ *partitionings* of Φ. These two partitionings are $\Theta_A = \{\theta_{A_1}, \ldots, \theta_{A_M}\}$ and $\Theta_B = \{\theta_{B_1}, \ldots, \theta_{B_M}\}$. For each pair of images I_i and I_j, there are four possibilities:

$$\begin{gathered}
((I_i \in \theta_{A_k}) \wedge (I_j \in \theta_{A_k})) \wedge ((I_i \in \theta_{B_\gamma}) \wedge (I_j \in \theta_{B_\gamma})) \\
((I_i \in \theta_{A_k}) \wedge (I_j \notin \theta_{A_k})) \wedge ((I_i \in \theta_{B_\gamma}) \wedge (I_j \notin \theta_{B_\gamma})) \\
((I_i \in \theta_{A_k}) \wedge (I_j \in \theta_{A_k})) \wedge ((I_i \in \theta_{B_\gamma}) \wedge (I_j \notin \theta_{B_\gamma})) \\
((I_i \in \theta_{A_k}) \wedge (I_j \notin \theta_{A_k})) \wedge ((I_i \in \theta_{B_\gamma}) \wedge (I_j \in \theta_{B_\gamma})).
\end{gathered}$$

The first two equations describe cases in which subjects A and B have agreed that images I_i and I_j are either similar or dissimilar, and the second two describe cases of disagreement. We define a binary variable $X_{ij}(\Theta_A, \Theta_B)$, which is 1 when A and B agree about i and j, and 0 otherwise. A *raw agreement measure* can be obtained by counting the number of cases of agreement. A *normalized agreement measure*, S, where $S = 0$ indicates complete disagreement and $S = 1$ complete agreement, can then be defined as

$$S(\Theta_A, \Theta_B) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X_{ij}(\Theta_A, \Theta_B) \tag{1}$$

This normalized agreement measure has a problem: it fails to take into account chance agreements, which has been shown to be extremely important [24, 25]. A more appropriate agreement measure is Cohen's kappa statistic [25]:

$$\begin{aligned}
\kappa(\Theta_A, \Theta_B) &= \frac{\text{observed agreement} - \text{expected chance agreement}}{1 - \text{expected chance agreement}} \\
&= \frac{S(\Theta_A, \Theta_B) - E[S(\Theta_A, \Theta_B)]}{1 - E[S(\Theta_A, \Theta_B)]}.
\end{aligned} \tag{2}$$

The value of $E[S]$ depends upon subject behaviour. When subsets are labeled, a Bayesian approach to its estimation is usually adopted. We have shown that assuming that subjects assign images to subsets with equal probabilities is inadequate, and derived a means of extending the Bayesian approach to the case of unlabeled subsets. We call the agreement statistic derived in this way κ_B; it ranges from $\frac{-E[S(\Theta_A,\Theta_B)]}{1-E[S(\Theta_A,\Theta_B)]}$ to 1. In practice, however, only the positive part of its range is used: we can usually design a system which does better than chance! Details of its derivation may be found in [22].

3.2 Agreements between and amongst humans and machines

We used κ_B to measure the agreement between the partitionings of a set of 100 images into 8 subsets by a group of 18 human subjects. The experiments showed that there was great variation in the partitionings produced by the human subjects, but the agreement between subjects was always significantly greater than that expected by chance. The average value κ_B between all pairs of human subjects was 0.3450. The maximum and minimum values were 0.6266 and 0.1736 respectively. These numbers might be thought of as a benchmark for the performance that could be expected from a machine image partitioning system on this task.

18 varieties of factor-analysis-based image classification systems were applied to the same set of images. The average agreement between machine and human partitionings was 0.1067, and the extreme values were 0.0250 and 0.2312. Clearly, these machine techniques failed to capture the common component of human image similarity judgment. We propose to use machine learning to seek a better result.

4 Frequency-based similarity

We would like to use the ground-truth data provided by human image partitionings to improve the performance of machine image partitioning techniques. We thus need a way of converting the human partitionings into similarity-based distances between pairs of images, since some distance forms the basis of most clustering techniques used to produce partitionings of an image set.

4.1 Distance definition based on human similarity judgments

We propose a distance based on the frequency with which human subjects judge a pair of images to be dissimilar. If all subjects place a pair of images in the same subset, the distance between them is 0. If all subjects place a pair of images in different subsets, the distance between them is 1. Let the distance between images I_i and I_j be $d_f(I_i, I_j)$. For P subjects, let $k \in [1, \binom{P}{2}]$ index each possible pair of subjects A_k and B_k. Then

$$d_f(I_i, I_j) = \frac{\sum_{k=1}^{\binom{P}{2}} 1 - X_{ij}(\Theta_{A_k}, \Theta_{B_k})}{\binom{P}{2}}. \tag{3}$$

A matrix of these distances was calculated for the 100 images partitioned by the 18 users in the experiment described above. It is difficult to visualize these 4950 distances so we present an image of the matrix of $d_f(I_i, I_j)$ values in Figure 1.

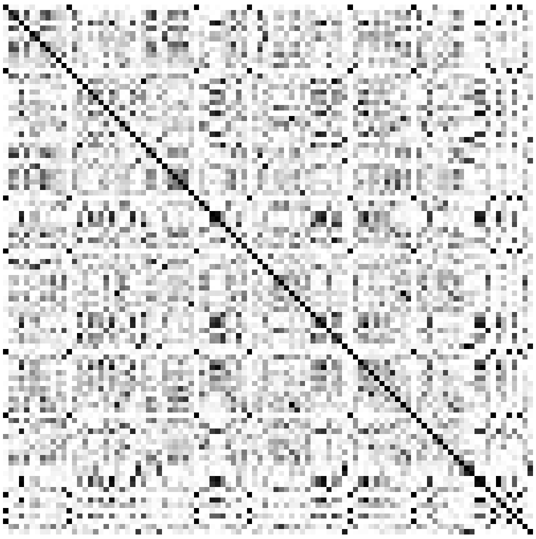


Figure 1: Image similarity distances calculated from human partitionings of the image set.

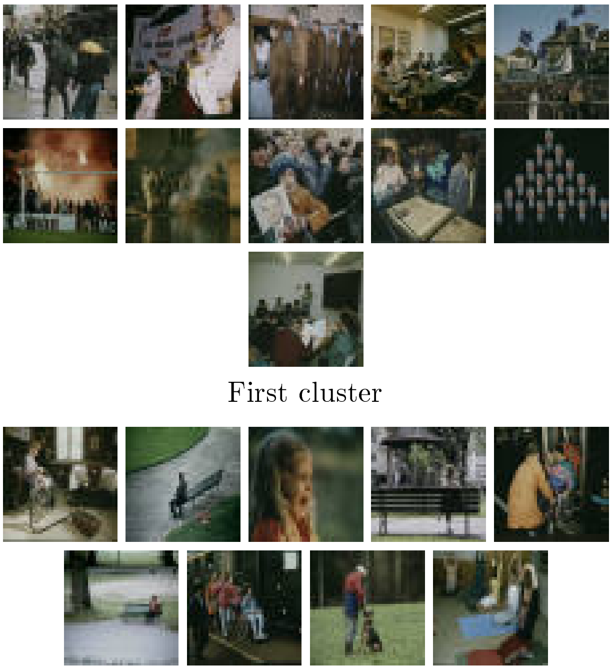
The black diagonal line corresponds to the distance between an image and itself: it is always zero. Further interpretation is possible. The white lines, with black points at their intersections, correspond to images of bank-notes, which all subjects placed in a single class. Consequently, they are maximally dissimilar (white lines) to every thing except each other (black points).

4.2 Clustering from a distance matrix

Since these distances are not derived from image locations in a feature space, geometric clustering techniques, such as Ascendant Hierarchical Clustering used in the earlier study

[22], could not be applied, as one cannot calculate the distances between clusters based on the coordinates of their centres. Images and clusters simply do not have coordinates.

The Unweighted Pair Group Method [26] was applied to cluster the images based on the distance matrix. The closest pair of images or clusters is found by exhaustive search, and these are merged to form a new cluster. There is a number of ways of computing the distance between this new cluster and the other images or clusters. One could, for example, take the arithmetic mean of the distances between the merged clusters and the others. Several techniques were tried, and the best results were obtained using simply the sum of the distances to the other clusters. Two of the eight clusters from the 3rd level of the resultant binary tree are shown in Figure 2.



First cluster

Second cluster

Figure 2: Two of the eight clusters from the 3rd level of the binary tree.

These might look like "good" clusters: the first could perhaps be labeled "crowd scenes", and the second possibly "people outdoors". We have seen in our earlier experiments, however, that this kind of subjective assessment of the "goodness" of a partitioning is dangerous. Humans are extremely good at finding explanations for why a group of images belong together. In order to perform a rigorous assessment of the "goodness" of this clustering, we use our agreement measure, κ_B. The agreements between this machine clustering and the 18 human clusters are shown in Table 1.

0.4458	0.3331	0.2837	0.3706	0.4174	0.4121
0.3371	0.4149	0.3246	0.4823	0.4350	0.4056
0.4724	0.4532	0.4686	0.3814	0.4852	0.3800

Table 1: Agreements between the frequency-based similarity clustering and human partitionings calculated using κ_B.

The average agreement was $\mu_{\kappa_B} = 0.4057$. Remarkably, this is greater than the average agreement between the human clusterings (0.3450) used to derive the distance matrix. This clustering agrees quite well with all the human clusterings. This result indicates that this “frequency of dissimilarity”-based distance is a good candidate for the common factor in human judgments of image similarity. The problem now, is how to use it.

5 Generalizing this distance

If ground truth data, in the form of human partitionings, were available for all the images in a database, this measure could be used directly. This, however, is unlikely ever to be the case. We would like to relate this distance measure to numerical image features, so that a distance could be calculated between images which have never been seen by a human user. We seek, in effect, to learn a mapping from feature space to perceptual similarity space.

The universal approximation property of neural networks makes them good candidates for learning such a mapping. We have applied multilayer perceptrons, trained by back-propagation, to the task. The target output is the similarity between a pair of images, as determined by Equation 3, when the input consists of numerical features extracted from the images. The networks are of the form shown in Figure 3.

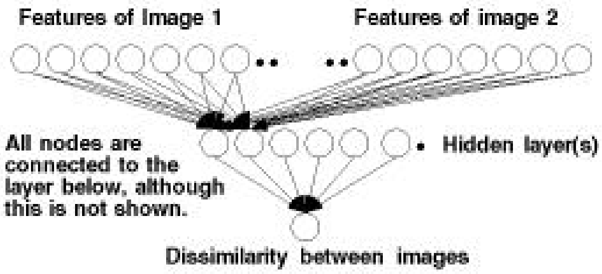
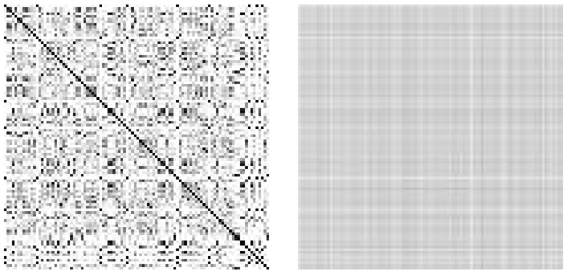

Features of Image 1
Features of image 2
All nodes are connected to the layer below, although this is not shown.
Hidden layer(s)
Dissimilarity between images

Figure 3: Form of networks used for mapping image features to image distance.

5.1 Selected networks applied to learning image similarity

As a benchmark, we commenced with a network with no hidden layer: a linear system. Figure 4 shows the distance matrix output by the network, and the target matrix. It is clear that the network has failed to capture most of the structure of the target similarity function.



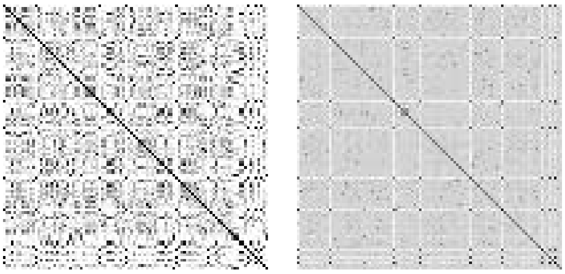
(a) target (b) network output

Figure 4: Failure of linear system to capture similarity matrix structure.

Such a network is only useful if it can be applied to images not used during training. To assess this aspect of network performance, the frequency-based distances derived from a second human image clustering experiment were used as a test set during the training of the networks. The error on the test data diverged: not only did the linear network fail to fit the training data well, it also failed to generalize.

The average agreement between the clustering based on the output of the linear network and the human clusterings was 0.0707. In earlier experiments, we found that the average agreement between factor analysis-based clusterings and the human clusterings was 0.1067, and the extreme values were 0.0250 and 0.2312 [22, 23]. The linear network's performance is actually not so bad by comparison.

A variety of more complex networks has also been tried. For example, a network with two hidden layers, each with 16 nodes, had the performance shown in Figure 5.



(a) target (b) network output

Figure 5: Performance of a network with two hidden layers, each with 16 nodes.

Due to the size of the training set, these networks are very slow to train. The results above were obtained after 5 days of training on a SPARC station 5. Importantly, both training and test set errors were still decreasing at that stage. The average agreement between the clustering produced by this network and the human clusterings is 0.1586, already better than the average of the factor analysis-based techniques, and could be expected to improve as training continues.

The process of selecting and evaluating various architectures and training strategies is on-going. Although these results are promising, it seems clear that a better set of features is required. Since the number of images for which such human similarity judgments can be obtained will always be relatively small, some form of regularization will also be required, to prevent the network from overfitting the data. This, and improved learning techniques, are the immediate future areas of research.

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] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The QBIC system," *IEEE Computer*, pp. 23–32, September 1995.

[4] "Excalibur Visual RetrievalWare SDK 2.1 Technical summary." web page, 1997.

[5] 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.

[6] T. Kato, "Database architecture for content-based image retrieval," *SPIE Image Storage and Retrieval Systems*, vol. 1662, pp. 112–123, 1992.

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

[8] F. Mokhtarian and S. A. J. Kittler, "Efficient and robust retrieval by shape content through curvature scale space," in Smeulders and Jain [27], pp. 35–42.

[9] A. Vellaikal and C.-C. J. Kuo, "Content-based image retrieval using multiresolution histogram representation," in *Digital Image Storage and Archiving Systems* (C.-C. J. Kuo, ed.), no. 2606 in SPIE Proceedings, (Philadelphia, PA, USA), pp. 312–323, October 1995.

[10] T. P. Minka and R. W. Picard, "Interactive learning using a "society of models"," in CVPR'96 [28], pp. 447–452.

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

[12] W. Ma and B. Manjunath, "Texture features and learning similarity," in CVPR'96 [28], pp. 425–430.

[13] F. Liu and R. Picard, "Periodicity, directionality, and randomness: Wold features for image modeling and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 722–733, July 1996.

[14] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Tools for content-based manipulation of image databases," *International Journal of Computer Vision*, vol. 13, June 1996.

[15] R. Zarita and S. Lelandais, "Wavelets and high order statistics for texture classification," in Frydrych *et al.* [29], pp. 95–102.

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

[17] A. D. Bimbo and P. Pala, "Shape indexing by multi-scale representation," in Smeulders and Jain [27], pp. 43–50.

[18] G. P. Robinson, H. D. Targare, J. S. Duncan, and C. C. Jaffe, "Medical image collection indexing: Shape-based retrieval using KD-trees," *Computerized Medical Imaging and Graphics*, vol. 20, no. 4, pp. 209–217, 1996.

[19] S. D. Cohen and L. J. Guibas, “Shape-based image retrieval using geometric hashing,” in *Proceedings of the ARPA Image Understanding Workshop*, May 1997.

[20] M. Cherbuliez, “Wavelet-based image querying,” Tech. Rep. 97.02, Computer Vision Group, Computing Centre, University of Geneva, rue Général Dufour, 24, CH-1211 Genève, Switzerland, February 1997.

[21] A. Tversky, “Features of similarity,” *Psychological Review*, vol. 84, pp. 327–352, July 1977.

[22] D. M. Squire and T. Pun, “A comparison of human and machine assessments of image similarity for the organization of image databases,” in Frydrych *et al.* [29], pp. 51–58.

[23] D. M. Squire and T. Pun, “Assessing agreement between human and machine clusterings of image databases,” Tech. Rep. 97.03, Computer Vision Group, Computing Centre, University of Geneva, rue Général Dufour, 24, CH-1211 Genève, Switzerland, April 1997. (submitted for puiblication).

[24] J. J. Bartko and W. T. Carpenter, “On the methods and theory of reliability,” *The Journal of Nervous and Mental Disease*, vol. 163, no. 5, pp. 307–317, 1976.

[25] 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.

[26] P. Sneath and R. Sokal, *Numerical Taxonomy: The Principles and Practice of Numerical Classification.* San Francisco: W. H. Freeman and Company, 1973.

[27] A. W. M. Smeulders and R. Jain, eds., *Image Databases and Multi-Media Search*, (Kruislaan 403, 1098 SJ Amsterdam, The Netherlands), Intelligent Sensory Information Systems, Faculty of Mathematics, Computer Science, Physics and Astronomy, Amsterdam University Press, August 1996.

[28] *Proceedings of the 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96)*, (San Francisco, California), IEEE Computer Society Press, June 1996.

[29] M. Frydrych, J. Parkkinen, and A. Visa, eds., *The 10th Scandinavian Conference on Image Analysis (SCIA'97)*, (Lappeenranta, Finland), Pattern Recognition Society of Finland, June 1997.