A Comparison of Human and Machine Assessments of Image Similarity for the Organization of Image Databases

Scandinavian Conference on Image Analysis June 9–11, 1997, Lappeenranta, Finland

David McG. Squire, Thierry Pun
Computer Science Department, University of Geneva
24, rue Général Dufour
CH-1211 Geneva 4, Switzerland

Abstract

There has recently been much interest in the organization and *content-based* querying of large images databases. Most frequently, the underlying hypothesis is that image similarity can be characterized by low-level image features, without further abstraction. This assumes that there is sufficient agreement between machine and human measures of image similarity for the database to be useful. We wish to assess the veracity of this assumption. To this end, we develop measures of the agreement between two partitionings of an image set; we show that it is vital to take chance agreements into account. We then use these measures to assess the agreement between human subjects and a variety of machine clustering techniques on a set of images. The results can be used to select and refine image distance measures for querying and organizing image databases.

Keywords: image similarity, image database organization, agreement statistics.

1 Introduction

There has recently been significant growth in research into methods for organizing and querying large databases of images of unconstrained real-world scenes. In particular, means of querying such databases by *image content* have been sought, rather than by using text labels attached by human experts [7, 3, 9, 11, 6]. The use of query-by-content stems from the belief that simple text labels are too terse to describe image contents adequately. A measure of image similarity is required, both for evaluating the similarity of between stored images and the query image, and for organizing the database so that this search may be performed efficiently.

In this context, it is largely acknowledged that the general object recognition problem remains unsolved [3]. Consequently, much work on content-based queries uses images from restricted domains, such as industrial trademarks [7, 6], or marine animals [10]. Even then, a semantic description is not attempted. Rather, a variety of low-level image features has been used, such as colour histograms [3], segment and arc statistics [12, 9], or shape measures [10, 13]. In all these cases, the hypothesis is that image similarity can be characterized by combinations of these low-level features, without moving to a higher level of abstraction. It is assumed that there is sufficient agreement between machine and human measures of image similarity for the database to be useful to all. In this paper we will assess the veracity of this assumption.

In section 2 we discuss the need for a measure of image similarity, propose an experiment for assessing human judgment of image similarity, and consider how such data could be used to rate and improve machine measures. In section 3 we define a measure of the agreement between two partitionings of an image set into unlabeled subsets, and develop statistics which indicate the degree to which the measured agreement is *better than that expected by chance*. The method for assessing the agreement between human and machine partitionings of an image set appears in section 4. The results and implications of these experiments are discussed in section 5.

2 Image similarity

Any image database system needs a measure of *image similarity*. The Euclidean distance between points in a multidimensional feature space is often used. Whatever measure is chosen, it is a function of some image features.

In short, the designer tries to select image features and a distance function so that the resultant distance is a measure of image similarity.

The aim of an image database system is to assist a human user to retrieve images. In systems which use query by image content, the query itself is an image. The system computes the distance between the query and the stored images and returns those images which are "close". This implies that the system's measure of image distance corresponds to the user's notion of the dissimilarity between images. Moreover, most such systems do not adapt to individual users, implying that the notion of image similarity is shared. The assumption is that there is sufficient overlap between the machine and human measures of image similarity for the database to be useful to all.

In this paper we attempt to assess the validity of this assumption. Our approach is to get human subjects to divide a set of images into a number of subsets, with no guidance. A measure of the agreement between pairs of subjects is derived in section 3.1, based on statistical reliability measures [1, 2]. The intent is to measure the degree of consistency of human image similarity measures. This permit an assessment to be made of how reasonable it is to expect *any* machine measure of image similarity to agree with a human user.

There are many image features to choose from when constructing a measure of image distance. A technique for reducing the dimensionality of the feature space, and a distance measure must also be selected. Finally, there are many clustering techniques from which to choose when organizing the database.

It has been extremely difficult to make an objective assessment of the performance of such systems, and thus to compare them, due to the lack of large sets of images for which the "ground truth" is known. In contrast, much text retrieval research uses data from the same large, expert-classified datasets. Quantitative comparisons between systems are made, notably in the TREC conferences.[1] Some attempts have been made to use human subjects' performance ratings to choose between and optimize the vision algorithms. An example, using edge detectors, is found in [5].

We use a variety of machine systems to cluster the same images that are presented to the human subjects. We then compute the agreement between the machine and the human partitionings. Averaging allows us to measure the agreement between machine measures of image similarity and the common human measure. Thus the machine systems can be ranked, and a choice made.

3 Statistical measures of agreement

3.1 Definition of the agreement measure

We consider subjects who partition a set of N images into M *unlabeled* subsets. We wish to measure the *agreement* between subjects: the similarity of their image similarity measures. We estimate this by considering their agreement on pairs of images. There is a literature on such comparisons, notably in medicine and psychology, where the typical aim is to measure agreement between physicians' diagnoses [1, 2]. Our problem differs from that paradigm since, in our case, the subsets are unlabeled.

We define an *agreement measure* based on pairs of images. Consider a set of images $\Phi = \{I_1, \ldots, I_N\}$. Subjects A and B independently partition Φ into M subsets. The resultant *partitionings* of Φ are $\Theta_A = \{\theta_{A_1}, \ldots, \theta_{A_M}\}$ and $\Theta_B = \{\theta_{B_1}, \ldots, \theta_{B_M}\}$. For images I_i and I_j, there are four possibilities:

$$((I_i \in \theta_{A_k}) \wedge (I_j \in \theta_{A_k})) \wedge ((I_i \in \theta_{B_\gamma}) \wedge (I_j \in \theta_{B_\gamma})). \quad (1)$$

$$((I_i \in \theta_{A_k}) \wedge (I_j \notin \theta_{A_k})) \wedge ((I_i \in \theta_{B_\gamma}) \wedge (I_j \notin \theta_{B_\gamma})). \quad (2)$$

$$((I_i \in \theta_{A_k}) \wedge (I_j \in \theta_{A_k})) \wedge ((I_i \in \theta_{B_\gamma}) \wedge (I_j \notin \theta_{B_\gamma})). \quad (3)$$

$$((I_i \in \theta_{A_k}) \wedge (I_j \notin \theta_{A_k})) \wedge ((I_i \in \theta_{B_\gamma}) \wedge (I_j \in \theta_{B_\gamma})). \quad (4)$$

In cases 1 and 2, A and B agree that I_i and I_j are either similar or dissimilar; in cases 3 and 4 they disagree. We define a variable $X_{ij}(\Theta_A, \Theta_B)$ which is 1 when A and B agree about images i and j, and 0 otherwise. A *raw agreement measure*, $S_{\text{raw}}(\Theta_A, \Theta_B)$, is obtained by counting agreements:

$$S_{\text{raw}}(\Theta_A, \Theta_B) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X_{ij}(\Theta_A, \Theta_B). \quad (5)$$

[1] **T**ext **RE**trieval **C**onference – further information is available at: `http://potomac.ncsl.nist.gov/TREC/`.

A *normalized agreement measure* is defined in Equation 6, ranging from 0 (total disagreement) to 1 (total agreement):

$$S(\Theta_A, \Theta_B) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X_{ij}(\Theta_A, \Theta_B). \tag{6}$$

As an example, for the first two humans studied we obtained $S(\Theta_1, \Theta_2) = 0.8123$ (see section 5.1). At first glance, this seems to be a high level of agreement. This is misleading. We have not considered chance agreements. Random partitionings of the image set will have some agreement, arising purely by chance. If we model the partitioning process, we can compute the expected value of this chance agreement, and thus correct the agreement measure. Failure to consider chance agreement is a common failing in the image database literature, even though its necessity is well-known in other fields. Typically, if any comparison is made, a raw intersection of machine and user responses to a query image is given, usually for a small image database (eg. [4]).

A better agreement measure is Cohen's κ statistic [2]:

$$\kappa(\Theta_A, \Theta_B) = \frac{\text{observed agreement} - \text{expected chance agreement}}{1 - \text{expected chance agreement}} = \frac{S(\Theta_A, \Theta_B) - E[S(\Theta_A, \Theta_B)]}{1 - E[S(\Theta_A, \Theta_B)]}. \tag{7}$$

The true value of $E[S]$ depends upon subject behaviour. In the absence of such information, we can estimate $E[S]$ by assuming that a "blindfolded" subject assigns images to each subset with equal probability. $E[S]$ is the expected value of S for two such subjects. X_{ij} is now a random variable, and we can find its distribution. We will call this the *blindfolded subject model* (BSM).

Consider a pair of images, I_i and I_j. Each image is a member of a subset in the partitionings Θ_A and Θ_B. For image I_i, let us label these subsets: $I_i \in \theta_{A_k}$, and $I_i \in \theta_{B_\gamma}$. We then have:

$$\begin{aligned} \Pr(I_j \in \theta_{A_k}) &= \frac{1}{M} & \Pr(I_j \in \theta_{B_\gamma}) &= \frac{1}{M} \\ \Pr(I_j \notin \theta_{A_k}) &= \frac{M-1}{M} & \Pr(I_j \notin \theta_{B_\gamma}) &= \frac{M-1}{M}. \end{aligned} \tag{8}$$

Using Equations 1 through 4, we obtain

$$\Pr(X_{ij} = 1) = \frac{1 + (M-1)^2}{M^2} \qquad \Pr(X_{ij} = 0) = \frac{2(M-1)}{M^2}. \tag{9}$$

We can now calculate the statistics of S, since it is the result of $\binom{N}{2}$ trials of a binomial process:

$$E[S] = \frac{1 + (M-1)^2}{M^2} \qquad \sigma^2[S] = \frac{4(M-1)(1 + (M-1)^2)}{N(N-1)M^4}. \tag{10}$$

We can use these in turn to compute the statistics of κ:

$$E[\kappa] = 0 \qquad \sigma^2[\kappa] = \frac{1 + (M-1)^2}{N(N-1)(M-1)}. \tag{11}$$

In the experiment described here, $N = 100$ and $M = 8$, giving $E[S] = 0.7813$ and $\sigma[\kappa] = 0.0269$. For the subjects mentioned above, we obtain $\kappa(\Theta_1, \Theta_2) = (0.8123 - 0.7813)/(1 - 0.7813) = 0.1420$. Their agreement, which at first seemed very high, is only 14.20% better than that expected by chance. It is, however, more than 5 standard deviations from the mean, so it is still significant.

3.2 An alternative agreement measure

The BSM has short-comings, arising from the assumption that images are assigned to subsets with equal probabilities. The problem becomes obvious if one considers two blindfolded subjects, A and B, who each assign all images to a single subset. We obtain $\kappa_{AB} = 1$, indicating 100% improvement over chance. However, if these

subjects always behave like this, then $\kappa(AB) = 1$ *always*. The apparent "perfect agreement" between A and B arises entirely from their bias, and conveys no information about their judgments of image similarity.

Preliminary experiments indicated that both human and computer partitionings produced subsets of greatly varying sizes (see section 5.2 for details). The existence of this bias means that the BSM will not account well for the observed data. This problem can be avoided by taking into account the frequency with which subjects assign images to subsets. This is can be understood by considering a matrix $\mathbf{A}$ of intersections between subsets. Table 1(a) shows an hypothetical example with $N = 20$ and $M = 3$. The row and column sums from Table 1(a) are

Table 1: Intersections between subsets created by subjects A and B.

	θ_{A_1}	θ_{A_2}	θ_{A_3}	a_{+j}
θ_{B_1}	8	2	1	11
θ_{B_2}	2	3	0	5
θ_{B_3}	2	1	1	4
a_{i+}	12	6	2	20

(a) Observed

	θ_{A_1}	θ_{A_2}	θ_{A_3}
θ_{B_1}	6.6	3.3	1.1
θ_{B_2}	3.0	1.5	0.5
θ_{B_3}	2.4	1.2	0.4

(b) Expected

used to compute "Bayesian" expected values for the intersections, using Equation 12 [2]. The matrix of expected intersections is shown in Table 1(b).

$$E_B[a_{ij}] = \frac{a_{i+}a_{+j}}{N}. \tag{12}$$

$S_{\text{raw}}(AB)$ (see Equation 5) can be computed from Table 1(a),

$$S_{\text{raw}}(AB) = \sum_{1 \le i,j \le M} \binom{a_{ij}}{2} + \frac{1}{2} \sum_{\substack{1 \le k,l,m,n \le M \\ k \ne m,\, l \ne n}} a_{kl} a_{mn}, \tag{13}$$

where the usual definitions $\binom{1}{2} \equiv 0$ and $\binom{0}{2} \equiv 0$ are made. The first sum in Equation 13 counts agreements of the form specified by Equation 1, and the second counts agreements corresponding to Equation 2. The factor of $1/2$ arises because each pair of elements appears twice.

This allows us to obtain an approximation to $E[S_{\text{raw}}]$. An exact calculation requires the enumeration of all possible assignments of images to subsets, since S_{raw} is a nonlinear function of the $\{a_{ij}\}$. Rather than performing this time-consuming calculation for each pair of partitionings considered, we will approximate $E[S_{\text{raw}}]$ by $\tilde{E}[S_{\text{raw}}] = S_{\text{raw}}(\{E[a_{ij}]\})$, the value of S_{raw} computed from the matrix of expected intersections. The values must be scaled to integers so that $\binom{a_{ij}}{2}$ can be computed.[2] The value obtained is then normalized to obtain $\tilde{E}_B[S]$, which is used to calculate a new statistic, κ_B, based on this new estimate of the expected chance agreement.

For Table 1(b), $\tilde{E}_B[S] = 0.5063$. The BSM gives $E[S] = 0.5556$. The agreement measures are $\kappa = -0.0066$ and $\kappa_B = 0.0937$. Taking into account the subjects' biases yields a κ_B indicating better than chance agreement, whereas the BSM gives a result worse than chance. The disadvantage of κ_B is that the expected agreement is subject-pair-dependent, making the interpretation of data more difficult.

4 Experiments

Human subjects were asked to partition a set of N images into at most M subsets. This was done using a program running on SUN workstations, which presented each subject with a source image set, and M 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.

For this experiment, the source image set consisted of 100 colour images selected at random from a set of 500 unconstrained images provided by Télévision Suisse Romande (a subset of the full 10,000 images provided). These images contained some "runs" of images from the same video footage, thus some highly similar images could be expected. The images were randomly ordered so that similar images would not necessarily be presented adjacently. Sample images are shown in Figure 1.

[2] Alternatively, the extension of the binomial coefficient function in terms of gamma functions could be used.

Figure 1: Sample images (originals were in colour).

Subjects were given a demonstration, and told that the notion of image similarity was entirely their choice. The task was performed by 10 members of the computer vision research group at the Université de Genève (who may be considered to be have some expert knowledge), and by 8 undergraduate students and lay-people.

The images were also classified into a binary tree using Ascendant Hierarchical Classification (AHC), using a variety of distance measures derived by applying Correspondence Analysis (CA) [8], Principal Components Analysis (PCA) and Normalized Principal Components Analysis (NPCA) to a range of colour, segment and arc statistics [12, 9]. For each factor analysis technique, the classification was performed both with and without the inclusion if colour features. For each case, a classification was done using 2, 4 and all of the ranked factors. The third level of a binary tree contains at most 8 classes, and is thus comparable with the human classifications described above.

5 Results and discussion

5.1 Agreement between humans

Table 2 shows a summary of the agreement between pairs of human subjects, as provided by κ. The average agreement between human subjects is 18.53% better than that expected from the BSM. Whilst perhaps lower than expected, this is more than 6 standard deviations from the mean. In fact, using the BSM, $\Pr(\kappa \geq 0.1853) = 9.202 \times 10^{-13}$. The "blindfolded' subjects hypothesis may be rejected.

Table 2: Statistics summarizing the agreement between human subjects, using κ.

	mean	median	std. dev.	κ min.	κ max.
All Subjects	0.1853	0.1919	0.1241	-0.1627	0.4708
Experts	0.2368	0.2335	0.0972	0.0294	0.4708
Lay People	0.1359	0.1300	0.1604	-0.1627	0.4246

These data also show that the agreement between experts is significantly higher than that between lay people, and that the variation between experts is less than that between lay people. This result indicates that the human image similarity measure may be partially learnt.

As an aside, the subject with the lowest average agreement with other subjects was colour-blind. This hints that colour information is important in human judgments of image similarity, which is not unexpected.

Table 3 shows a summary of the κ_B values for the human subjects. The trends remain the same. The mean values are higher, and the overall variance is lower, than for the κ. This is expected, as κ_B explicitly models observed subject biases.

5.2 Agreement between machine partitionings

The average κ for machine techniques is -0.2662: *worse* than if the machines merely assigned images to subsets with uniform probability. This seems a depressing result for proponents of these image clustering methods. The behaviour of these techniques, however, does not correspond to the BSM. All the machine clusterings contained

Table 3: Statistics summarizing the agreement between human subjects, using κ_B.

	mean	median	std. dev.	κ_B min.	κ_B max.
All Subjects	0.3450	0.3381	0.0926	0.1736	0.6266
Experts	0.3773	0.3625	0.0822	0.2225	0.5868
Lay People	0.3181	0.2781	0.1128	0.1736	0.6266

some very large subsets, and some very small ones. This is because the distribution of factor values is usually bell-shaped. This results, using AHC, in several large subsets close to the mean values of the factors, and small subsets containing the remaining images. The mean maximum subset size was 34, with standard deviation 8.4. The mean minimum subset size was 1.7 (1.2). For the human subjects, the mean maximum was 24 (4.7), and the mean minimum was 4.5 (2.0). This suggests that a clustering technique which produced subsets of more uniform size than AHC would better match human performance.

This bias means that κ_B is preferable. Table 4 summarizes κ_B for the machine partitionings. Taking the non-uniform subsets sizes into account, the agreement between all techniques is better than that expected by chance. CA has the greatest self-agreement under changes of the number of factors retained and NPCA the least, but the values are within one standard deviation of each other. Little conclusion can be drawn about the relative stability of these techniques.

Table 4: Statistics summarizing the agreement between machine partitionings, using κ_B.

	mean	median	std. dev.	κ_B min.	κ_B max.
All Variations	0.2023	0.1574	0.1757	0.0176	0.9743
All CA	0.3404	0.0848	0.3437	0.0675	0.9743
CA colour	0.6185	0.4508	0.2518	0.4303	0.9743
CA no colour	0.8364	0.8134	0.0605	0.7765	0.9192
All NPCA	0.2862	0.2537	0.0949	0.2026	0.5511
NPCA colour	0.3019	0.2569	0.1037	0.2036	0.4453
NPCA no colour	0.4116	0.3518	0.0990	0.3319	0.5511
All PCA	0.3144	0.0877	0.2987	0.0626	0.8014
PCA colour	0.7824	0.7780	0.0141	0.7677	0.8014
PCA no colour	0.5517	0.5202	0.0673	0.4897	0.6453

A self-agreement of 0.3404 (CA) seems very low. From the raw data, it could be seen that the self-agreement of these techniques is much higher when considering only the retention of 4 factors against all factors. For CA, it is 0.9743 when colour features are used, and 0.9192 otherwise. This confirms that the data is well represented by only the first four factors (the first four factors explain 88.35% of the variance in the CA space when colour features are included).

5.3 Agreement between human and machine partitionings

It was established in section 5.2 that the bias of human and machine partitionings towards large subsets makes κ_B the better agreement measure. It is still interesting to note some observations using κ. The mean κ between machine and human partitionings was -0.2806. The best technique, averaged over the six variants, was PCA, with -0.1844. The worst was CA, with -0.4663. Averaged over different numbers of retained factors, the best was PCA with colour features, with $\mu_{\mathrm{PCA_c}}[\kappa] = -0.1393$, and the worst was CA with no colour, with $\mu_{\mathrm{CA_{nc}}}[\kappa] = -0.5060$. The overall standard deviation was 0.1648, so these results are significant. The *only* human subject with any positive κ was subject H_{16}: the first-named author of this paper, who has worked with these machine image clustering methods for 12 months. This result points to the danger of allowing the creators of image similarity measures to be their sole assessors. Secondly, and positively, it seems to indicate that human measures of image similarity are partially learnt: after prolonged interaction with an image database system, the human begins to judge image similarity in a similar way.

Table 5 summarizes κ_B for the human and machine partitionings. Using κ_B, the agreement between machine and human partitionings is positive in all cases. The machine technique with the least improvement over chance

Table 5: Summary of agreement between human and machine partitionings, using κ_B.

	mean	median	std. dev.	κ_B min.	κ_B max.
All Variations	0.1067	0.1058	0.0338	0.0250	0.2312
All CA	0.0915	0.0836	0.0371	0.0250	0.1825
CA colour	0.1169	0.1186	0.0308	0.0683	0.1825
CA no colour	0.0662	0.0621	0.0226	0.0250	0.1229
All NPCA	0.1233	0.1235	0.0281	0.0589	0.1911
NPCA colour	0.1388	0.1400	0.0235	0.0817	0.1911
NPCA no colour	0.1078	0.1068	0.0233	0.0589	0.1633
All PCA	0.1052	0.1007	0.0276	0.0553	0.2312
PCA colour	0.1167	0.1086	0.0296	0.0719	0.2312
PCA no colour	0.0936	0.0914	0.0197	0.0553	0.1328

was CA, no colour features, all factors retained, with $\mu[\kappa_B] = 0.0573$. The best was NPCA, colour features, all factors retained, with $\mu[\kappa_B] = 0.1477$. The standard deviation over all machine clustering variants was 0.0240, so this is significant. When averaged over all its variants, NPCA was the best technique, with 0.1233, and CA was worst, with 0.0915, with a standard deviation of 0.0130. Again, this would seem to be significant. For each factor analysis technique, the use of colour features gave improved agreement with the human subjects, corroborating the conjecture in section 5.1.

Over all techniques, the greatest κ_B with any subject was 0.2312 (again with the first-named author). This compares with 0.6266 between H_4 and H_{18}. The average κ_B between humans and humans was 0.3464, whereas between humans and machine partitionings it was 0.1067. The machine techniques reviewed here do provide significantly better than chance agreement with human subjects, but are a long way from being as good as the "average" human.

6 Conclusion

We have shown that a rigorous assessment of the agreement between two partitionings of a set of images into unlabeled subsets is possible. Such a measure *must* take account of the expected chance agreement. Expected chance agreement depends on the model of user behaviour selected. The blindfolded subject model does not describe the measured behaviour of human subjects or machine partitioning techniques. We have thus proposed the κ_B statistic, which takes into account the subset probabilities for each subject.

Using κ_B, we found that the average agreement between humans is 34.64% higher than that expected by chance, almost 4 standard deviations away from the expected value. This indicates that there is some shared notion of image similarity, but it is far from 100% agreement. There is also much variance between pairs of humans: different subjects agree in different ways. It is likely that the appropriate measure will depend not only on the user, but also on the genre of images and the task being performed. This suggests that an image database system should model the user, so that the image distance measure is partially learnt. There is evidence that the human notion of image similarity is partially learnt. It seems that humans can adapt to an image classification system. This suggests that users will be able to adapt to a image database system, as well as *vice versa*.

Agreement between human and machine partitionings was much less than that between pairs of humans. Despite this, the human partitionings provide an objective means of assessing image clustering systems. It was possible to conclude that colour features should be used, and that Normalized Principal Components Analysis was the best factor analysis technique tried. More importantly, the utility of the methodology has been demonstrated. Larger scale experiments would be necessary to obtain clearly significant distinctions between competing measures of image similarity.

Finally, we reiterate our belief in the importance of gathering ground truth for assessing the performance of image database systems. We have demonstrated a methodology for analyzing and applying such data, and, in particular, shown the importance of using statistics which consider chance agreements.

Acknowledgments

This work is supported by the Swiss National Foundation for Scientific Research (grant no. 2100-045581.95). The authors would like to thank Marianne Gex-Fabry for her advice on measures of agreement and reliability.

References

[1] J. J. Bartko and W. T. Carpenter. On the methods and theory of reliability. *The Journal of Nervous and Mental Disease*, 163(5):307–317, 1976.

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

[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*, pages 23–32, September 1995.

[4] K. Han and S.-H. Myaeng. Image organization and retrieval with automatically constructed feature vectors. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson, editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 157–165, Zürich, Switzerland, August 1996.

[5] M. Heath, S. Sarkar, T. Sanocki, and K. Bowyer. Comparison of edge detectors: A methodology and initial study. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 143–148, San Francisco, California, June 1996. IEEE Computer Society Press.

[6] A. K. Jain and A. Vailaya. Image retrieval using color and shape. *Pattern Recognition*, 29(8):1233–1244, August 1996.

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

[8] L. Lebart, A. Morineau, and J.-P. Fénelon. *Traitement des données statistiques; méthodes et programmes.* Dunod, Paris, 1979.

[9] R. Milanese, D. Squire, and T. Pun. Correspondence analysis and hierarchical indexing for content-based image retrieval. In P. Delogne, editor, *IEEE International Conference on Image Processing*, volume III, pages 859–862, Lausanne, Switzerland, September 1996.

[10] F. Mokhtarian and S. A. J. Kittler. Efficient and robust retrieval by shape content through curvature scale space. In Smeulders and Jain [14], pages 35–42.

[11] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Tools for content-based manipulation of image databases. *International Journal of Computer Vision*, 13(3), June 1996.

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

[13] S. Sclaroff. Encoding deformable shape categories for efficient content-based search. In Smeulders and Jain [14], pages 107–114.

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