In Fourth IEEE Workshop on Applications of Computer Vision (WACV'98), Princeton, NJ, USA, October 1998, pp. 88–93.

Learning a similarity-based distance measure for image database organization from human partitionings of an image set

David McG. Squire*
University of Geneva
Computer Vision Group
rue Général Dufour, 24, CH 1211 Genève 4, Switzerland
David.Squire@cui.unige.ch

Abstract

In this paper we employ human judgments of image similarity to improve the organization of an image database. We first derive a statistic, κ_B which measures the agreement between two partitionings of an image set. κ_B is used to assess agreement both amongst and between human and machine partitionings. This provides a rigorous means of choosing between competing image database organization systems, and of assessing the performance of such systems with respect to human judgments.

Human partitionings of an image set are used to define an similarity value based on the frequency with which images are judged to be similar. When this measure is used to partition an image set using a clustering technique, the resultant partitioning agrees better with human partitionings than any of the feature-space-based techniques investigated.

Finally, we investigate the use multilayer perceptrons and a Distance Learning Network *to learn a mapping from feature space to this perceptual similarity space. The Distance Learning Network is shown to learn a mapping which results in partitionings in excellent agreement with those produced by human subjects.*

1. Introduction

The rapid growth of the world wide web and the use of digital images in the preparation of paper documents mean that millions of people now access multimedia documents daily. Multimedia documents contain images, either static or as video frames. There is thus a need for systems that allow users to create, manage and query image databases in an efficient and accurate manner. The attachment of text labels to images is inadequate, since identical images can be described in different ways, and controlled vocabulary indexing is now deemed insufficient even in text retrieval systems. Consequently, there is 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 [6, 25]. There are now several commercial CBIRSs available, such as IBM's QBIC [5] and the Virage system [6]. 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. We seek to improve the performance of CBIRSs by using machine learning to incorporate *human similarity judgments* in the process of database organization. Resultant systems should have better measures of image similarity than those based solely upon image features.

We have performed experiments to measure the agreement between human partitionings of an image set, as well as agreement between human and machine partitionings. We have developed a *measure of the agreement* between two such partitionings, based on pair-wise subset membership comparisons. Random partitionings can have significant chance agreement. We have derived a better, *chance-corrected*, agreement measure. The expected chance agreement can be large, especially for the small image sets often used to test CBIRSs. It is *vital* to take this into account.

Agreement between humans is significantly better than chance, but much less than might have been anticipated. Agreement between human and machine partitionings is not as great. No single similarity measure can be expected to satisfy all users.

We envisage a complete CBIRS architecture which exhibits a gradual transition from an expert-designed feature space to a user- and task-specific "query-interaction space" (See Figure 1). In this paper, we are concerned with the

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

third stage: the "shared similarity space". Although a complete system will develop individual user models, we see a role for an initial mapping from feature space to a space in which distances reflect image similarities commonly perceived by humans.

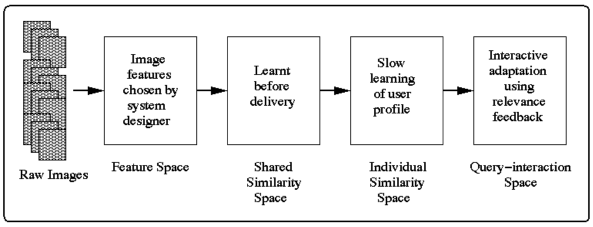


Figure 1. Architecture of proposed complete CBIRS

We show how human partitionings of an image set can be used to define a similarity value for each pair of images. This value leads to partitionings which agree better with human partitionings than any other method tried. Finally, we demonstrate a system which learns a mapping between image features and this similarity space.

2. State of the art

2.1. Features

Semantic retrieval remains impossible; *e.g.* 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 working in restricted domains, such as industrial trademarks [7] or marine animals [13]. In general, an attempt is made to capture similarity using some function of a set of low-level image features.

The most frequently used feature is colour [7, 18]. Similarity is defined as some distance between colour distributions, most commonly the colour histogram [23, 6]. Many systems use texture features, such as hierarchies of Gabor filters [10], the Wold features [9] used in Photobook [14], coarseness, contrast, and directionality in QBIC [5], or wavelet-based decompositions [24]. Importantly, images may have similar global colour or texture statistics, but little visual similarity, due to differing spatial distributions of these features.

Shape features are also often global (one shape per image), and are thus best applied to restricted domains. Modal matching has been applied to fish, rabbits and machine tools [17]. Other shape-based approaches include multi-scale representation of curves [2], histograms of edge directions [7, 16] and maxima of zero-crossing contours of curvature scale space images [13].

Global descriptors can be augmented by features which retain spatial information, such as Daubechies' or Haar wavelet decompositions [25]. Alternatively, images may be segmented into regions, from which features are extracted, such as colour, size, location and relationships to other regions. This approach adds labeled graph matching to the image retrieval problem.

2.2. Similarity

CBIRSs aim to return images which, according to human perception, are *similar* to a query image. Remarkably, few such systems consider what similarity means in the context of human usage. Those that do report that human similarity judgments similarity noticeably differ (*e.g.* [13]). 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 are *similar* to the query.

It is often implied that given the "right" features (an appropriate colour space [23, 18], texture features "corresponding to human perception" [9]), proximity in feature space *must* correspond to perceptual similarity. There are several reasons to doubt this. Most fundamentally, there is psychophysical evidence that human similarity judgments do not obey the requirements of a metric: self-identity, symmetry and the triangle inequality [22].

Some authors have addressed this problem. Self-organizing maps have been used to cluster texture features according to class labels provided by human judgments [10]. Minka and Picard report a system which learns groupings of similar images from positive and negative examples provided by users during query sessions [11, 12]. Their approach is very similar in spirit to the present work, although the set-based learning methods applied differ from the direct mapping from feature space to similarity space presented here. The approach we discuss avoids the need to recompute groupings whenever a new image is added to the dataset.

3. Image similarity and agreement between partitionings of a set

It is difficult to assess objectively 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, which permits the quantitative comparison of document retrieval systems.

In order to investigate human similarity judgments, we asked human subjects to partition a set of unconstrained colour images 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 [21], based on statistical measures of reliability well-known in medical and psychological research [1, 4].

The manner in which a user partitions an image set depends, of course, on the task which the user is performing. We chose not to specify any task or criteria in advance precisely because this is the implicit assumption made by CBIRSs which include neither learning nor relevance feedback. It is with respect to this baseline that we wish to compare the various systems' performances.

We used a variety of machine systems to cluster the same set of images. The agreement between the machine and the human partitionings was computed. Averaged over all humans, this provides a measure of the overlap of each machine measure of image similarity with the common human measure, which can be used to rank competing systems. 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

In measuring the agreement between two partitionings of an image set, pairs of images are considered individually (since the subsets are unlabeled). Consider the set of images $\Phi = \{I_1, \dots, I_N\}$. Two subjects, A and B, independently partition Φ into M subsets. The resultant *partitionings* of Φ are $\Theta_A = \{\theta_{A_1}, \dots, \theta_{A_M}\}$ and $\Theta_B = \{\theta_{B_1}, \dots, \theta_{B_M}\}$. For each pair of images I_i and I_j, there are four possibilities:

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

In the first two cases subjects A and B agree that images I_i and I_j are either similar or dissimilar, and in the second two they disagree. 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 *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 measure has a problem: it fails to correct for chance agreements, which has been shown to be extremely important [1, 4]. A better agreement measure is Cohen's kappa statistic [4]:

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

$E[S]$ depends on subject behaviour. We have shown that assuming that subjects assign images to subsets with equal probabilities is inadequate, and derived a means of extending the usual Bayesian approach to the case of unlabeled subsets [21]. The resultant statistic, κ_B, ranges from $\frac{-E[S(\Theta_A, \Theta_B)]}{1-E[S(\Theta_A, \Theta_B)]}$ to 1. In practice only the positive part of its range is used: we can usually design a system which does better than chance!

3.2. Agreement between and amongst humans and machines

We used κ_B to measure the agreement between partitionings of 100 images[1] into at most 8 subsets by a group of 18 human subjects (a maximum of 8 subsets was chosen since powers of two correspond to the maximum number of subsets at each level of a binary tree, and to facilitate the simultaneous viewing of all subsets). As might be expected for such an unconstrained task, there was great variation between the partitionings produced, but κ_B was always significantly greater than zero. The average κ_B between all pairs of human subjects was 0.3450. The maximum and minimum values were 0.6266 and 0.1736. 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 [21]. The average agreement between machine and human partitionings was 0.1067. 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 want to use the ground-truth data provided by human image partitionings to improve the performance of machine image set 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 partitioning techniques.

[1]Images were selected at random from a set of 500 unconstrained images provided by Télévision Suisse Romande.

We propose a distance based on the frequency with which human subjects judge a pair of images to be dissimilar. 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, B_k).

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

Since these distances are not derived from locations in a feature space, geometric clustering techniques can not be applied, since distances between clusters based on their centre coordinates cannot be computed. Images and clusters simply do not have coordinates.

The Unweighted Pair Group Method [20] was applied to cluster the images based on the distance matrix defined by Equation 3. 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, such as the arithmetic mean of the distances between the merged clusters and the others. Several techniques were tried, and the best results, as measured using κ_B, were obtained using the sum of the distances to the other clusters. 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.

The average agreement was 0.4057. Remarkably, this is greater than the average agreement between the human clusterings used to derive the distance matrix. This suggests that this "frequency of dissimilarity"-based distance is a good candidate for the common factor in human judgments of image similarity.

5. Generalizing this distance

If ground truth data were available for all images in a database, this measure could be used directly. This, however, is unlikely. We want to relate this measure to image features, so that distances can be calculated between images never seen by a user. We seek a mapping from feature space to perceptual similarity space.

5.1. Multilayer perceptrons

Multilayer perceptrons, trained by backpropagation, were applied to the task. The target output was the similarity between a pair of images (Equation 3). The input consisted of colour, segment, arc and region features extracted from the images.

A variety of networks was tried. The average agreement between the clustering produced by a network with two 16 node hidden layers and the human clusterings was 0.1586. In earlier experiments, the average agreement between factor analysis-based clusterings and the human clusterings was 0.1067 [21]. This is thus an improvement. Increasing the dimensionality of the network produced little change, suggesting that the features used do not contain enough information for the desired mapping to be learnt.

5.2. Distance-learning networks

A new class of self-organizing network, the distance-learning network (DLN), has been developed and applied to this task. Based on the self-organizing feature maps (SOMs) introduced by Kohonen [8], the DLN differs from the standard SOM in several ways. First, nodes have both input and output vectors. Standard SOMs have only input vectors, though nets with output vectors have been used previously, *e.g.* in robot control [15]. These vectors allow an input map and an output map simultaneously. These maps may have differing dimensionalities, but neighbourhood relationships in both are determined by the network topology. A manifold of the dimensionality of the network is thus embedded in both the input and output spaces.

The most significant difference between a DLN and a SOM is the learning rules. At each iteration, *two* input vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, are presented. The nodes having the closest input weights $\mathbf{w}_i$ to the input vectors, n_1 and n_2, are found. The $\mathbf{w}_i$ are updated according to

$$\mathbf{w}_i^{t+1} = \mathbf{w}_i^t + \epsilon \left[e^{-\frac{d_{1i}^2}{2\sigma}} (\mathbf{v}_1 - \mathbf{w}_i^t) + e^{-\frac{d_{2i}^2}{2\sigma}} (\mathbf{v}_2 - \mathbf{w}_i^t) \right], \quad (4)$$

where t is the timestep, ϵ is a scale factor, d_{ij} is the distance between nodes n_i and n_j *in the network topology* and σ is a radius of influence. ϵ and σ decrease as a function of t. This is just the vector sum of two normal SOM update steps, and the behaviour of the input mapping is exactly that of a SOM.

In the output space only the desired distances between activated nodes are given. If the distance between the output vectors of nodes n_i and n_j is greater than that desired, they attract $(+)$, otherwise, they repel $(-)$. Neighbours are affected as above. The update rule for the output weights $\mathbf{o}_i$

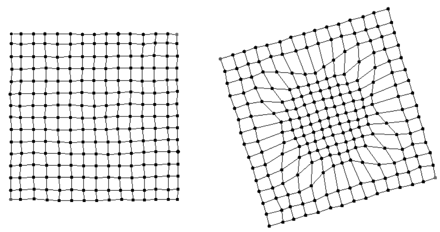
(attraction) is

$$\mathbf{o}_i^{t+1} = \mathbf{o}_i^t \pm \epsilon \left[e^{-\frac{d_{1i}^2}{2\sigma}} (\mathbf{o}_2^t - \mathbf{o}_i^t) + e^{-\frac{d_{2i}^2}{2\sigma}} (\mathbf{o}_1^t - \mathbf{o}_i^t) \right]. \quad (5)$$

The input map learnt reflects the frequency distribution and topology of the input vectors. If the dimensionality of the network is less than that of the input subspace, the network manifold will "fold itself" into it in a manner analogous to a generalized nonlinear PCA [15]. The DLN allows a distorted version of this topology to be learnt as the output of the network.

In a CBIRS using well-chosen features, the *topology* of feature space should be meaningful, even if absolute distances are not. Metaphorically, if two images are similar, we would like to drag them closer together in similarity space. If topology is meaningful, they should drag their neighbours with them. The DLN realizes this goal. The influence on neighbours is controlled by σ.

Figure 2 shows how a distorted output space can be learnt by a DLN, whilst preserving a topology determined by the input space. The network, its inputs and its outputs were all two-dimensional. Input vectors were distributed uniformly in the unit square. The target output distance was $\sqrt{|\mathbf{v}_1 - \mathbf{v}_2|^2}$, except when $\mathbf{v}_1$ and $\mathbf{v}_2$ fell in a circle of radius $1/\sqrt{8}$ centred at the origin, where it was halved.



(a) input map (b) output map

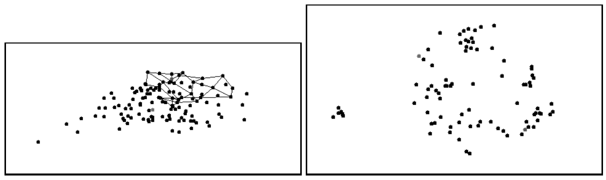
Figure 2. Input and output of a DLN. The centroid and orientation of the output map are arbitrary.

A variety of DLNs of different architectures were applied to the feature $\rightarrow$ similarity mapping task. We report results for 5 three-dimensional networks of $5 \times 5 \times 5$ nodes, with 16-dimensional input vectors and three-dimensional output vectors, trained with pairs of images drawn from the set of 100 used by the human subjects. Networks were assessed using the average agreement between partitionings resulting from clustering based on the output distances with all human subjects. The results appear in Table 2. We recall that the average intra-human agreement was 0.3450. The fourth column of Table 2 shows network performance as a percentage of this benchmark value. The DLNs do capture the common component of human similarity judgments for these images.

	mean	std. dev.	% of avg. hum.
Network 1	0.3413	0.0600	98.93
Network 2	0.3205	0.0583	92.90
Network 3	0.3469	0.0627	100.6
Network 4	0.3556	0.0461	103.1
Network 5	0.3300	0.0453	95.65

Table 2. Agreements between clusterings based on DLN similarity clusterings and human partitionings.

Figure 3 shows the input and output maps of Network 5 projected onto the first two dimensions of the input and output spaces. The clusters in the output map are readily apparent. Another advantage of the frequency-based similarity distance is that clusters between which there was confusion become neighbours in the DLN output space, since their members have similarity values less than one with each other. This means that a nearest-neighbour search should retrieve relevant images even when the radius extends beyond a given cluster.



(a) input map (b) output map

Figure 3. Input and output maps of Network 5

It should be noted that the values of image features play no part in this mapping. Only their topological relationships have a role. The output map is thus a form of topological look-up table, all though 4950 distances have been encoded (approximately) using only 375 variables.

6. Conclusion

We have proposed a measure of the agreement between two partitionings of an image set. This measure, κ_B, has the advantage that it is a point measure. Also, since similarity judgments about all images in the dataset are obtained, effectively, simultaneously during the partitioning process, obtaining data for calculating κ_B should be cheaper in user hours than gathering relevance judgments for a set of queries. This is in contrast to the precision/recall graphs often used to assess CBIRS performance. We believe that

κ_B complements the precision/recall approach, particularly for evaluating systems which use clustering to organize the database for faster search.

We have shown how human partitionings of an image set can be used to define a frequency-based similarity measure which leads to partitionings in excellent agreement with those produced by human subjects. We have introduced a new class of self-organizing network, the *Distance-Learning Network*. We have demonstrated that DLNs can learn a mapping from feature space to similarity space using the frequency-based similarity measure as a target during training. Partitionings of images sets obtained by clustering in the this learnt similarity space were in excellent agreement with human subjects, the average being 98.24% of the mean intra-human agreement.

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] A. D. Bimbo and P. Pala. Shape indexing by multi-scale representation. In Smeulders and Jain [19], pages 43–50.

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

[4] 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.

[5] 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*, 28(9):23–32, September 1995.

[6] A. Gupta and R. Jain. Visual information retrieval. *Communications of the ACM*, 40(5):70–79, May 1997.

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

[8] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[9] F. Liu and R. Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):722–733, July 1996.

[10] W. Ma and B. Manjunath. Texture features and learning similarity. In CVPR'96 [3], pages 425–430.

[11] T. Minka. An image database browser that learns from user interaction. Master's thesis, MIT Media Laboratory, 20 Ames St., Cambridge, MA 02139, 1996.

[12] T. P. Minka and R. W. Picard. Interactive learning using a "society of models". In CVPR'96 [3], pages 447–452.

[13] F. Mokhtarian and S. A. J. Kittler. Efficient and robust retrieval by shape content through curvature scale space. In Smeulders and Jain [19], pages 35–42.

[14] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Tools for content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3):233–254, June 1996.

[15] H. Ritter, T. Martinez, and K. Schulten. *Neural computation and self-organizing maps: an introduction*. Computation and neural systems series. Addison-Wesley Publishing Company, 1992.

[16] 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*, 20(4):209–217, 1996.

[17] S. Sclaroff. Deformable prototypes for encoding shape categories in image databases. *Pattern Recognition*, 30(4):627–642, April 1997. (special issue on image databases).

[18] 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.

[19] 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.

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

[21] D. M. Squire and T. Pun. Assessing agreement between human and machine clusterings of image databases. *Pattern Recognition*, 31(12), 1998. (to appear).

[22] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, July 1977.

[23] A. Vellaikal and C.-C. J. Kuo. Content-based image retrieval using multiresolution histogram representation. In C.-C. J. Kuo, editor, *Digital Image Storage and Archiving Systems*, volume 2606 of *SPIE Proceedings*, pages 312–323, Philadelphia, PA, USA, October 1995.

[24] R. Zarita and S. Lelandais. Wavelets and high order statistics for texture classification. In M. Frydrych, J. Parkkinen, and A. Visa, editors, *The 10th Scandinavian Conference on Image Analysis*, pages 95–102, Lappeenranta, Finland, June 1997. Pattern Recognition Society of Finland.

[25] 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*, pages 13–24, Washington D.C., May 1997.