An Image Database Semantically Structured based on Automatic Image Annotation for Content-Based Image Retrieval

Xuejian Xiong[†], Kap Luk Chan, Lei Wang
School of Electrical and Electronic Engineering
Nanyang Technological University
Nanyang Avenue, Singapore 639798
[†]Email: P144845039@ntu.edu.sg

Abstract

In this paper, we presented a semantically structured image database for content-based image retrieval. A class descriptor is proposed to represent each class using a multi-prototype model, which can be obtained by using a learning scheme, such as the Unsupervised Optimal Fuzzy Clustering algorithm, on a group of sample images manually selected from the class. Based on the proposed Image-Class Matching Distance, a similarity measure at the semantic level between an image and classes, images can be annotated by tokens of classes. Hence, composite features of images, including low-level descriptors, class descriptors, and image annotation, are stored into a structured feature database corresponding to the semantically structured image database. From experiments, it can be concluded that the performance of the semantically structured CBIR system is improved greatly in terms of retrieval time and efficiency.

1. Introduction

Effective image indexing and retrieval techniques are important and critical to facilitate people searching information from large image databases. In recent years, there are intensive research activities in Content-Based Image Retrieval (CBIR) systems[8, 7, 5, 15, 4, 10, 6, 3, 1, 11]. In these systems, images are represented using a set of low-level descriptors, i.e. colour, texture, shape, etc. Such descriptions are also a focus of recent MPEG-7 standard committee[9]. However, there is the semantic gap between low-level descriptors of images and the meaningful interpretation of images by users. In order to introduce high-level information of images into a CBIR system, images are labeled using captions, text, keywords, etc., manually, such as what the MIT's Photobook[8] has done. Such a manual process is too labor intensive for annotating thousands images. Therefore, we try to annotate images automatically, and then organize these images into a semantically meaningful structure. The annotation aims to capture the major image content corresponding to human understanding. The problem of automatic image annotation is therefore the same as that of image classification. To solve it, there are two issues needed to be addressed, i.e. the representation of classes and the association of a new image with a certain class.

In [11], vacation image classes are described using class-conditional densities by studying low-level features of training images under the constraint that images belongs to one and only one class. Bayesian decision rule is adopted to classify the test image. In [1], images are described by the nodes of a weighted graph, and classification is based on nonmetric distances. Another method group images into meaningful categories based on low-level features using a self-organizing map[15].

In this paper, we propose the class descriptors which can capture high-level concepts, through a learning scheme, of images in a supervised fashion. First of all, images are stored in a semantically structured database, which means images are put into image classes defined manually and semantically based on human understanding. Note that an image can belong to two classes or more simultaneously. There can be subclasses if necessary. Because the high-level concept of an image class is very complex and highly nonlinear, we assume here that it can be modeled by multiple prototypes, which can be obtained by using a learning scheme on a group of sample images selected manually from a class. Each image in the database is categorized into a class based on a proposed similarity measure, i.e. Image-Class Matching Distance (ICMD), between the image and classes at the semantic level. ICMD can match an image with a class which is represented by a multi-prototype model. After image categorization, the image can then be annotated by the token of the class. Hence each image can now be described by a composite feature set that includes low-level descriptors, class descriptors, and image annotation. Composite features of images are then stored into a corresponding structured feature database. By introducing these new attributes into a CBIR system, the performance of image retrieval can be improved greatly. Motivated by the above considerations, a CBIR system with relevant feedback in a structured database, named CBIR_S system, is constructed.

The paper is organized as follows. The definition and generation of class descriptors are presented in Section 2, where Unsupervised Optimal Fuzzy Clustering (UOFC) algorithm used for constructing class descriptors is briefly described, and a similarity measure between an image and classes at the semantic level is proposed. The semantically structured image and feature databases of the CBIR_S system is presented in Section 3. Section 4 shows experiments done for evaluating image retrieval performance of the CBIR_S system. Finally, in Section 5, conclusions is given.

2. Class Descriptors Generation

A learning scheme is proposed to obtain high-level information of images in a supervised fashion. The bootstrapping of the learning scheme is the selection of some typical images, i.e. sample images, from each available class by human observers. Each class descriptor can then be generated by studying a set of low-level descriptors extracted from sample images of the class, i.e. sample features. This process mimics the learning process of humans. As a result, the system, just like an infant, can obtain the knowledge about an image, say a flower image, by studying some samples from the flower class. Then the concept of ‘Flower’ can be captured in the CBIR_S system quantitatively and a class descriptor can be constructed to represent the Flower class.

Because the description of an image class is a complex nonlinear problem, we assumed that it can be characterized by a multi-prototype model (see Figure 1). Therefore, the class l can be represented as $\{pc_1^l, pc_2^l, \cdots, pc_{C^l}^l\}$, where pc_j^l is the jth prototype in the class l, and C^l is the number of prototypes in the class l. In fact, the number of prototypes is not necessarily same in every class. Moreover, each prototype is not necessarily spherical and Gaussian. It can be in any arbitrary shape. In other words, each prototype can have any kind of distributions. In the case of Gaussian distributions, the class then can be represented by a Gaussian Mixture.

In order to find the multiple prototypes in a class, a learning scheme is proposed. Here Unsupervised Optimal Fuzzy Clustering (UOFC) algorithm is used. The UOFC algorithm performs automatic clustering of sample images of a class in a kind of feature spaces, and finds the optimal number of clusters. As a result, a cluster is a prototype of the class. Consequently, a set of parameters of multiple prototypes, e.g. centers, dispersion, size, etc., can be used to represent the class.

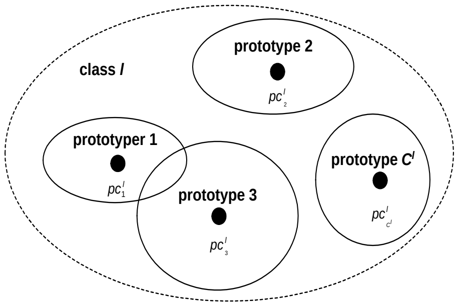


Figure 1. The multi-prototype model representing the class descriptor of the lth class.

2.1. The Unsupervised Optimal Fuzzy Clustering (UOFC) Algorithm

Let us consider a collection of p_l sample images in a class l, and each image is represented by a low-level descriptor vector $f_i^l, (i = 1, 2, \cdots, p_l)$, forming the input data set F^l. Then the new modified generalized objective function proposed based on [14] for UOFC algorithm is given as follows:

$$J(U,V;F) = \sum_{i=1}^{c} \sum_{k=1}^{p_l} (\mu_{ik})^m \{(1-g) * (\| f_k^l - V_i^l \|_p)^p + g * (\sum_{j=1}^{r} ((f_k^l - V_i^l) \bullet s_{ij})^2)\}, \tag{1}$$

where, c is the number of the clusters, symbol $(\bullet)$ denotes the inner product, $V_i^l, (i = 1, 2, \cdots, C^l)$ is the center of the ith cluster, $m \in [1, \infty)$ is a weighting exponent on each fuzzy membership, and $g \in [0, 1]$ is a weighted value. $U = \{\mu_{ik}\}$ is the fuzzy membership matrix, and μ_{ik} should satisfy the following two conditions:

(i) $\sum_{k=1}^{p_l} \mu_{ik} = 1$ for all k.

(ii) $p_l > \sum_{k=1}^{p_l} \mu_{ik} > 0$ for all i.

The details of the UOFC algorithm can be found in our previous work[13].

Note that $|| \cdot ||_p$ stands for p-norm distance measurement, where $p > 0$. Obviously, if $p = 2$, it is a Euclidean distance.

Then each cluster is spherical, and it can be described by a single Gaussian distribution. As a result, the class can be represented by a Gaussian Mixture. If $p = 1$, it is a Manhattan distance, and each cluster is rhomboidal. Moreover, if $p < 1$, it is a nonmetric distance[1].

In order to validate the UOFC algorithm, a similarity matrix $R = \{R_{ij}\}$ is used,

$$R_{ij} = \frac{dp_i + dp_j}{dv_{ij}}. \tag{2}$$

where, dp_i measures the dispersion of the ith cluster, and dv_{ij} describe the dissimilarity between the ith and jth cluster. As a result, the similarity criteria between two clusters is as follows,

1. The ith and jth cluster are separated, if $R_{ij} > 1$. Then the cluster number c is not changed.

2. The ith and jth cluster can be merged, if $R_{ij} \leq 1$. Then $c = c - 1$.

The advantage for using the similarity criteria is that less clustering iterations are needed to attain the optimal number of clusters, hence the computation complexity can be reduced greatly.

If one image class l can be clustered into C^l prototypes, and the jth prototype can be represented by the center V_j^l, the size w_j^l, and the dispersion dp_j^l, then the class l is described as

$$cf^l = \{w_1^l, V_1^l, dp_1^l, \cdots, w_{C^l}^l, V_{C^l}^l, dp_{C^l}^l\} \tag{3}$$

It is possible to sub-divided each class according to needs, and then subclass descriptors can be obtained in a similar way.

2.2. Similarity Measurement between Images and Classes at the Semantic Level

In order to categorize an image into a certain class represented by multi-prototype, more sophisticated similarity measurement between images and classes should be used.In [11], a Baysian framework is used to classify images based on obtained class-conditional densities. Unlike it, we proposed a Image-Class Matching Distance (ICMD), a modified Earth Mover's Distance (EMD)[10], to measure similarity between an image and a class at the semantic level. There are two main reasons for introducing the ICMD. One is that much research in psychology suggests that human similarity judgments are not metric[1], which means it does not obey the triangle inequality even though it is symmetric. Since classes are represented by the multi-prototype model, the similarity measurement between an image and a class is a matching problem between one point and several points in a kind of feature spaces. Obviously, the commonly used point to point distance calculation is not suitable.

Inspired by the Earth Mover's Distance (EMD), which has been successfully used for measuring image similarity[10, 3], Image-Class Matching Distance (ICMD) is proposed here to calculate the distance between a query image and a class at the semantic level. Let f_q be the feature of the image Q. The class descriptor of the lth class is $cf^l = \{pc_j^l\}$. Then the ICMD is defined as:

$$ICMD(Q, l) = \frac{\sum_{j=1}^{C^l} flow_{qj} dist(f_q, pc_j^l)}{\sum_{j=1}^{C^l} flow_{qj}} \tag{4}$$

where, w_j^l is the weight of the jth prototype in the lth class, and can be set as the percentage of the prototype's size n_j^l over the whole class N^l. $flow_{qj}$ is the optimal admissible flow from the query image Q to the class l that minimizes the numerator of Eqn. (4) subject to the following constraints:

$$\begin{aligned} & flow_{qj} \geq 0, \\ & \sum_{j=1}^{C^l} flow_{qj} \leq 1, \quad flow_{qj} \leq w_j^l \\ & \sum_{j=1}^{C^l} flow_{qj} = \min(1, \sum_{j=1}^{C^l} w_j^l). \end{aligned} \tag{5}$$

$dist(f_q, pc_j^l)$ is the ground distance. In this paper, we use the normalized Euclidean distance[5] as the ground distance. Obviously, the smaller the ICMD between the image and the class is, the more similar they are. If there is

$$c = \left\{ c | ICMD_{qc} = \min_{l=1}^{CS} \{ICMD_{ql}\} \right\} \tag{6}$$

Then it can be concluded that the image Q can be categorized to the cth class, and Q can be annotated by the token of the cth class, i.e. $\{a_c\}$.

3. The Semantically Structured Image and Feature Databases

In order to describe the characteristics of images, three kinds of descriptors consisting of the composite features are introduced. They are: 1) low-level descriptors, which represent characteristics of image contents. They can be global features and/or regional features, and can represent different kinds of attributes of image contents, e.g. colour, texture, shape, spatial, etc.; 2) class descriptors, which quantitatively represent features of image classes at the semantic level. Here, each class is described by a multi-prototype model; and 3) image annotation, using tokens of classes which images belong to. In fact, images are annotated using the semantically meaningful language. As a result, the composite feature set of an image I_j is

$$Feature_j = \{a_j, cf^l, f_j\} \tag{7}$$

where, a_j is the annotation of I_j, that is also the token of the class l. The image's low-level descriptor is represented by f_j. Therefore, composite features of images in the database are stored into a hierarchical composite feature database corresponding to the structured image database (see Figure 2). For a class l, based on the selection of a kind of low-level attributes of image contents, class tokens, class descriptors, and low-level descriptors are then constructed orderly to represent images in the class. The feature database corresponds to the image database, and both of them are hierarchically organized. Moreover, different kinds of low-level descriptors can be combined together to represent image contents more completely.

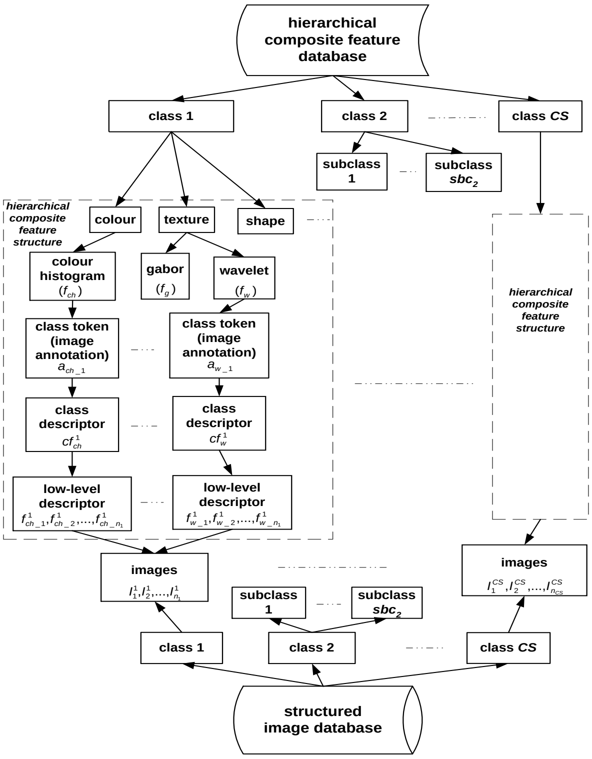

hierarchical composite feature database
class 1
class 2
class CS
subclass 1
subclass sbc_2
hierarchical composite feature structure
colour
texture
shape
colour histogram (f_{ch})
gabor (f_g)
wavelet (f_w)
class token (image annotation) a_{ch_1}
class token (image annotation) a_{w_1}
class descriptor cf^1_{ch}
class descriptor cf^1_w
low-level descriptor $f^1_{ch\_1}, f^1_{ch\_2}, \ldots, f^1_{ch\_n_1}$
low-level descriptor $f^1_{w\_1}, f^1_{w\_2}, \ldots, f^1_{w\_n_1}$
hierarchical composite feature structure
images $I^1_1, I^1_2, \ldots, I^1_{n_1}$
images $I^{CS}_1, I^{CS}_2, \ldots, I^{CS}_{n_{CS}}$
subclass 1
subclass sbc_2
class 1
class 2
class CS
structured image database

Figure 2. The structured composite feature database.

There are two parts in the CBIR system with relevance feedback in a structured database (CBIR_S), the off-line part and the on-line one. The main work in the off-line part is description of images, resulting in a hierarchical composite feature database. The image retrieval process is done on-line. When a query image is presented through the interaction scheme, it is processed to obtain the composite feature. Then a list of relevant classes can be obtained using ICMD. Only images within the most relevant class are compared with the query image. That means images in other classes are considered irrelevant, and they are not used at all. Therefore, the retrieval reports the first k most similar images, which is visualized on the user-machine interface. Based on the user's perception, relevance comments can be provided through relevance feedback, and then are introduced into the system to dynamically improve image retrieval results to satisfy the user's demand. Actually, users' feedback can also be returned into the generating process of image descriptors to revise representation of images.

4. Experimental Results

4.1. Performance Evaluation Measure

The retrieval efficiency measure used in [7] is adopted here as the performance criterion. For a query image q, by comparing with all K images (except the query image itself) using an index technique, the first $(N_q + T)$ images are retrieved. Here, T is a positive integer which is used as a tolerance to test the consistence of a retrieval algorithm. If n_q is the number of successfully retrieved similar images, the efficiency of retrieval can then be defined as:

$$\eta_R(T) = \left. \frac{\sum_{q=0}^{K} n_q}{\sum_{q=0}^{K} N_q} \right|_T \tag{8}$$

The evaluation of a CBIR system's performance looked at both retrieval time and efficiency. All images from the image database are used as query. When one image is selected as a query, this image is removed from the database.

4.2. Constructing a CBIR system with a flat database for Comparison

The CBIR system with a flat database (CBIR_F) means that all images are randomly put into the image database, and the corresponding low-level descriptors are consequently stored into a feature database randomly too. As a result, for retrieving similar images based on a selected query image, distance between the query image and each image in the image database is calculated. Here, the normalized Euclidean distance is used. Then the first k most similar images can be obtained.

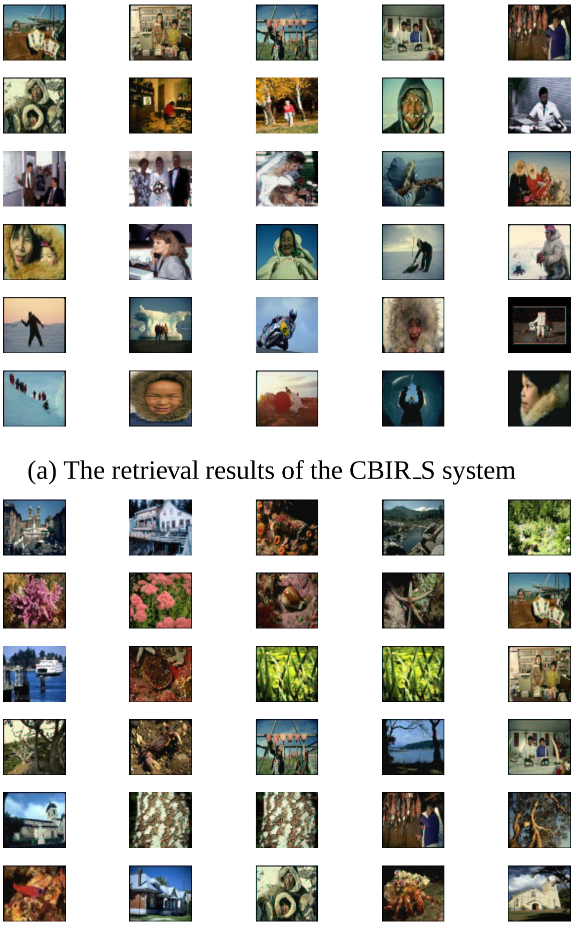
4.3. A General Image Database

1249 general images extracted from MIT VisTex and Corel Stock Photos are used here. All images are categorized into 14 classes by a group of human observers for the purpose of performance evaluation. Some classes include several subclasses. Sample images are selected for each class and subclass manually. In the present work, only colour and texture attributes are used to describe the contents of images. Therefore, there are three kinds of low-level descriptors of images, i.e. the global colour histogram feature $f_{ch} \in R^{256}$, the global gabor feature $f_g \in R^{48}$, and the global wavelet feature $f_w \in R^{24}$[5, 2, 12].

Table 1. The average retrieval time and efficiency ($T = 24$) of the CBIR_S and CBIR_F system, respectively.

Feat-	Average Retrieval Time (ms)		Average Efficiency (%)	
-ures	CBIR_S	CBIR_F	CBIR_S	CBIR_F
f_w	1.153	61.39	56.33	34.49
f_g	2.354	69.62	61.74	34.98
f_{ch}	3.371	171.52	58.30	30.72

From Table 1, it can be seen that the average retrieval time is reduced greatly in the CBIR_S system, say nearly one-sixtieth of that of the CBIR_F system. Also, the average retrieval efficiency is improved greatly. In Figure 3, the first 30 retrieval images are shown for using a human image as the query. The global gabor feature is used to represent image contents. It can be seen that, unlike the CBIR_F system, there is no irrelevant images in the retrieval results using the CBIR_S system.



(a) The retrieval results of the CBIR_S system

(b) The retrieval results of the CBIR_F system

Figure 3. The first 30 relevant images retrieved using the global gabor feature for a human image

The comparison of retrieval efficiency between the CBIR_F and CBIR_S system with different relevance feedbacks for using the global gabor features is show in Figure 4. It can be concluded that the retrieval efficiency of the CBIR_S system without relevance feedback is consistently and nearly 20% above that of the CBIR_F system. However, we should point out that image mis-annotation has adverse effects on the retrieval efficiency. But relevance feedback can compensate this error. Users can select from the n most relevant classes whether or not the automatic selection is wrong. Users can also remark the first presented retrieved image that whether it has the same annotation to the query image or not. If not, the system then presents retrieved images from the second most similar class or from user's selection. The procedure is iterated until the user obtains the best result. In fact, based on the structured database, it is easier and faster for a user to input his/her comments. Therefore, in Figure 4, after 3 relevance feedbacks, the retrieval efficiency of the CBIR_S system is over 80%. Moreover, in Figure 5, the comparison of retrieval efficiency for the global gabor feature by using relevance feedback between different values of T is given. It shows that T has no significant effect in retrieval results of the CBIR_S system. The reason is that as soon as the right class is retrieved, the retrieval result is the best already.

5. Conclusions

In this paper, we presented a semantically structured image database for content-based image retrieval. A class descriptor is proposed to represent each image class using a multi-prototype model, which can be obtained by using a learning scheme on a group of sample images manually selected from the class. The Unsupervised Optimal Fuzzy Clustering algorithm is used here to perform automatic clustering of sample features of the class, and finds the optimal number of clusters by grouping similar cluster together. Based on the newly proposed Image-Class Matching Dis-

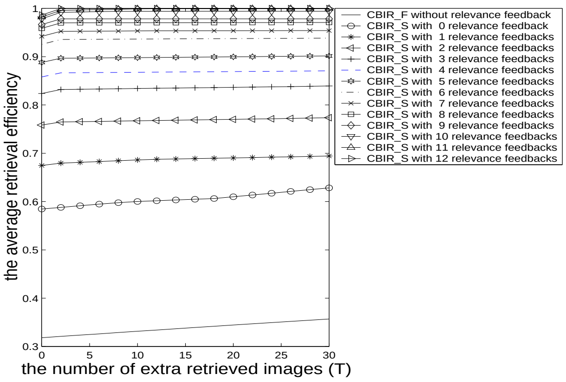


Figure 4. The comparison of retrieval efficiency between CBIR_S and CBIR_F with different relevance feedbacks for using the global gabor feature.

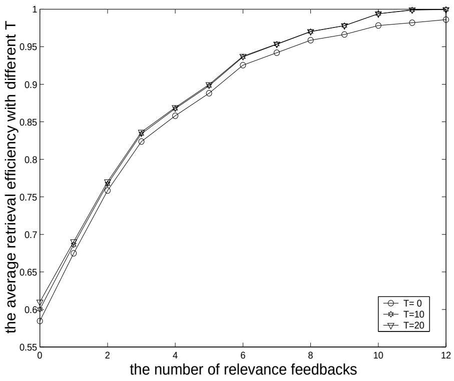


Figure 5. The comparison of retrieval efficiency for using the global gabor feature by relevance feedback between different values of T.

tance, a similarity measure at the semantic level between an image and classes, images can be annotated by tokens of classes. Hence, each image is described by a composite feature set that includes low-level descriptors, class descriptors, and image annotation. Composite features of images are then stored into a structured feature database corresponding to the semantically structured image database. The performance of the structure CBIR system is compared with that of a flat CBIR system by average retrieval time and efficiency. It can be seen from our experiments that the average retrieval time of the structure CBIR system is only one-sixtieth of that of the flat CBIR system, while the average retrieval efficiency is improved greatly.

References

[1] D.W.Jacobs, D.Weinshall, and Y. Gdalyahu. Classification with nonmetric distances: image retrieval and class representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(6):583–600, 2000.

[2] A. K. Jain and F. Farrokhnia. Unsupervised texture segmentation using Gabor filters. *Pattern Recognition*, 24(12), 1991.

[3] F. Liu, X. Xiong, and K. L. Chan. Natural image retrieval based on features of homogeneous color regions. In *Proceedings of the 4th IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 73–77, Austin, TX, U.S.A., April 2000.

[4] W. Y. Ma and B. S. Manjunath. Netra: a toolbox for navigating large image database. In *Proceedings of the International Conference on Image Processing*, volume 1, pages 568 –571, 1997.

[5] B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996.

[6] S. Medasani and R. Krishnapuram. A fuzzy approach to content-based image retrieval. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, pages 964–968, 1999.

[7] B. M. Mehtre, M. S. Kankanhalli, A. D. Narasimhalu, and G. C. Man. Color matching for image retrieval. *Pattern Recognition Letters*, 16, 1995.

[8] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Content-based manupulation of image databases. In *SPIE Storage and Retrieval of Image & Video Databases II*, pages 34–47, San Jose, CA, 1994.

[9] P.Salembier. Mpeg-7 multimedia description schemes. In *The 2nd MPEG-7 Awareness Event*, Singapore, March 2001.

[10] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Proceedings of the sixth IEEE International Conference on Computer Vision*, pages 59–66, 1998.

[11] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang. Image classification for content-based indexing. *IEEE Trans. on Image Processing*, 10(1):117–130, 2001.

[12] C. Wang. Content-based image indexing and retrieval: Feature extraction and feature similarity metrics. Master's thesis, Nanyang Technological University, 1998.

[13] X. Xiong and K. L. Chan. Towards an unsupervised optimal fuzzy clustering algorithm for image database organization. In *Proceedings of the 15th International Conference on Pattern Recognition*, pages 73–77, Barcelona, Spain, August 2000.

[14] Y. Yoshinari, W. Pedrycz, and K. Hirota. Construction of fuzzy models through clustering techniques. *Fuzzy Sets and Systems*, 54:157–165, 1993.

[15] D. Zhong, H. Zhang, and S.-F. Chang. Clustering methods for video browsing and annotation. In *Proceedings of the SPIE on Storage Retrieval Image Video Databases*, pages 239–246, San Jose, CA, U.S.A., 1996.