IMPROVING IMAGE RETRIEVAL WITH SEMANTIC CLASSIFICATION USING RELEVANCE FEEDBACK*

Hong Wu
National Lab of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
Beijing 100080, China
hwu@nlpr.ia.ac.cn

Mingjing Li, Hong-Jiang Zhang, Wei-Ying Ma
Microsoft Research Asia
49 Zhichun Road
Beijing 100080, China
{mjli, hjzhang, wyma}@microsoft.com

Abstract In this paper, we investigate the combination of image semantic classification with content-based image retrieval. A flexible scheme is proposed to take advantage of image classification, which may be obtained manually or automatically, to enhance image retrieval. In this scheme, a semantic feature vector is composed for an image based on its class membership information, and is combined with low-level features in image retrieval. Relevance feedback techniques are also used to adjust both the semantic feature and low-level features of the query in order to better reflect the user's intention. Experimental results on a collection of 10,000 images with manual classification demonstrate the effectiveness of the proposed method.

Keywords: content-based image retrieval, image classification, relevance feedback, semantic feature, and class membership.

* This work was performed at Microsoft Research Asia.

X. Zhou et al. (eds.), *Visual and Multimedia Information Management*

1. INTRODUCTION

The performance of most current content-based image retrieval (CBIR) systems (e.g. Pentland et al. 1994, Flickner et al. 1995, Smith & Chang 1996) is limited due to the gap between high-level concepts and low-level features, and the subjectivity of human visual perception. One approach to improving this situation is to employ relevance feedback (RF) techniques (Cox et al. 1996, Rui et al. 1998, Rui & Huang 2000). In the relevance feedback process, the system dynamically learns and refines the query and retrieval results to capture the user's intention. However, most of RF methods in CBIR systems are not so effective as in text information retrieval systems, due to the fact that refining low-level features to map semantic content is often difficult. Therefore, some works (e.g. La Cascia et al. 1998, Cox et al. 1997) engaged in incorporating semantic attributes with low-level features in image retrieval.

On the other hand, as image databases get large, a good organization is needed to aid better searching and browsing. Big image collections are often grouped into semantic classes, which may be further organized in a hierarchy. These semantic classes may be constructed manually, as did in most of the image search engines, or automatically using semantic attributes. Some efforts (e.g. Szummer & Picard 1998, Vailaya et al. 1998, Bradshaw 2000) attempted to use low-level features to automatically classify images into semantic classes, such as indoor, outdoor, man-made and natural, etc. WebSeek (Smith & Chang 1997) performs a semi-automatic classification of images on the Web into a hierarchy of subjects by using associated text and filename cues.

A straightforward way to combine image classification with CBIR is to use classes as a filter to restrict the search range, and to reduce both searching time and potential retrieval errors. For example, when performing image search in WebSeek, the user first selects a category through the hierarchy of subjects, browses and selects an image as the query example. Then, the searching process is constrained in that category. In Simplicity (Wang 2000), query image is first classified by the system as one of n predefined semantic classes (indoor-outdoor, objectionable-benign, and graph-photograph), and retrieval is enhanced by narrowing down the searching range to that particular category.

However, this method is not effective in some cases. (1) In some image databases, the number of image categories is large. Although they can be organized in a hierarchy, it is not easy for a common user to find the intended category. So users tend to perform retrievals at a high level in the hierarchy, and the detailed classification is not used. (2) When the user's need does not have apparent relations with these categories, due to the

subjectivity and diversity of human perception of image content, it is difficult for the user to make a choice. For example, there are some landscape images categorized as four classes, Asia, Europe, America, Africa, and Oceania. If a user want to find some images of beach, this categorization is not helpful to this query. And automatically constraining the searching range will not work in this case as well. Therefore, using image classes as a filter for image retrieval has a very limited success.

In this paper, a simple but flexible scheme is proposed to integrate semantic classification with low-level features for improving the performance of content-based image retrieval systems. Class membership of an image is used as the semantic feature and is represented as a Boolean vector. Such semantic feature and low-level features are used complementarily to retrieve images. And they are modified simultaneously through a relevance feedback process to better reflect user's perception of image content. This scheme may be considered a practical application of the hidden attribute approach described by Cox et al. (1997). Experimental result shows the effectiveness of the proposed method.

The rest of the paper is organized as follows. The proposed image retrieval scheme is presented in Section 2. The preliminary experimental results are presented in Section 3. Finally, we give discussion and concluding remarks in Section 4.

2. INCORPORATING CLASSIFICATION IN CBIR

Image classification can be regarded as treating a group of images as equivalent. Images within classes are more similar than those across classes in some way. The intuition of using semantic classification in CBIR is like following. If one or more images in a class are relevant to the user's query, other images in the same class are expected to be relevant in some extent as well.

To fully utilize the classification information existing in some image datasets, image classes are treated as pseudo keywords such that many techniques in text information retrieval (Baeza-Yates & Ribeiro 1999) may be deployed. For the sake of simplicity, a Boolean vector is formed for every image with each element corresponds to a class. If the image belongs to a particular class, the corresponding element in the vector is set to 1; it is set to 0 otherwise. This vector can be used as a semantic feature in addition to low-level features in image retrieval. Low-level features (e.g. Pentland et al. 1994, Flickner et al. 1995, Cox et al. 1996, Smith & Chang 1996) are visual statistics of images, and are characterized by feature vectors as well.

There are three main parts in the proposed scheme: semantic feature representation, integration in retrieval, and relevance feedback. We mainly present the details related to the semantic feature in this section.

2.1 Semantic Feature Representation

In our approach, image classes are treated as indexing terms in traditional information retrieval, and each term is assigned a binary weight. Term weight f_{ij} denotes the membership of image i to class j, i.e., $f_{ij}=1$ if $i \in j$, $f_{ij}=0$ if $i \notin j$. The semantic feature of image i is defined as an m-dimensional vector:

$$F_{ih}=(f_{i1}, f_{i2}, \cdots, f_{im})^T \tag{1}$$

where m is the total number of classes. In case that there is a class hierarchy, all high-level classes are also included.

As in text information retrieval, more sophisticated term weighting may be introduced in the feature representation, such as TF*IDF. Then, we get another definition:

$$F_{ih}'=(f_{i1}\times\log(\frac{N}{n_1}), f_{i2}\times\log(\frac{N}{n_2}), \cdots, f_{im}\times\log(\frac{N}{n_m}))^T \tag{2}$$

where N is the total number of images in the database, n_j is the number of images in class j. The new definition implies that classes with more images are less useful in retrieval.

2.2 Retrieval with Classification

In our scheme, the visual similarity and the semantic similarity between two images are calculated separately. The former is based on the low-level features, while the latter is based on the class membership. The overall similarity is the weighted sum of these two similarities, as defined in Equation (3):

$$S(q,i)=\omega * S_h(Q_h, F_{ih})+(1-\omega) * S_l(Q_l, F_{il}) \tag{3}$$

where q is the query and i is an image in the database, S_h and S_l are their semantic and visual similarities, respectively. Q_h and F_{ih} are semantic features of query and image i respectively, while Q_l and F_{il} are low-level features. ω is the semantic weight, subject to $0 \le \omega \le 1$.

The semantic similarity between the query and an image is calculated by the cosine of the angle between these two vectors:

$$S_h(Q_h, F_{ih}) = \frac{Q_h \bullet F_{ih}}{\|Q_h\| \cdot \|F_{ih}\|} \tag{4}$$

To compute the visual similarity, distance between the query and an image is first calculated by diagonally weighted Euclidean distance based on the low-level feature:

$$D(Q_l, F_{il}) = \sqrt{(F_{il} - Q_l)^T \Lambda (F_{il} - Q_l)} \tag{5}$$

$\Lambda = diag(w_1, ..., w_n)$, each diagonal element model the different importance of corresponding element of the low-level feature. The distance is further converted to visual similarity, which monotonously decreases as the distance increases. The similarity is normalized to the interval between 0 and 1.

2.3 Relevance Feedback

Relevance Feedback (RF) is an interactive learning technique, first introduced in text information retrieval. During retrieval, the user is asked to give feedbacks regarding the relevance of current outputs of the system, and the system learns from the feedbacks and refines the query to better capture the user's intention. In our scheme, different RF algorithms are adopted for semantic feature and low-level features respectively in the same relevance feedback process.

Initially, the semantic feature of the query is not available if the query example is new to the system, thus all elements in the vector are set to 0. On the other hand, if the example is from the database, its semantic feature vector may be used directly. However, it is not necessarily consistent with the user's intention. Therefore, relevance feedback is deployed to iteratively refine the semantic feature vector of the query in order to better reflect the user's intention.

Here, we adopt the Rocchio formula (Baeza-Yates & Ribeiro 1999) to refine the semantic feature of the query:

$$Q' = \alpha Q + \beta \frac{\sum Q^+}{n^+} - \gamma \frac{\sum Q^-}{n^-} \tag{6}$$

where Q is the original query, Q^+ is the set of positive (relevant) examples, n^+ is the number of positive examples, while Q^- and n^- are that of negative

(irrelevant) examples respectively, Q' is the updated query. This procedure automatically re-weights query terms by adding the weights from the actual occurrence of those query terms in the positive samples, and subtracting the weights of those terms occurring in negative samples. Query is also expanded using the term weights from the positive and negative samples. After updating, the semantic feature of query is no longer a Boolean vector. In our experiments, the query is treated as a positive example. So the parameters are set as $\alpha = 0$, $\beta = 1$, and $\gamma = 0.5$, to limit the effects of negative feedback. Actually, no negative weights are used in the updated query, i.e., a term weight is set to zero if it is negative. After relevance feedback, the class containing more positive examples and less negative examples will be assigned a larger weight, and more images from this class will be retrieved hopefully.

For relevance feedback on the low-level features, a simpler version of the general relevance feedback framework proposed by Rui & Huang (2000) was implemented in our system. The low-level feature vector of query (Q_l in Equation (5)) is updated to the average of features of all the relevant images. The weight of i-th element of low-level feature (w_i in Equation (5)) is set to inverse of σ_i^2, the variance of i-th element of all relevant low-level features.

3. EXPERIMENTAL RESULTS

In this section, we will compare the performance of the system adopting our scheme with that only based on low-level features. And further, to investigate the application of our scheme with different classification, we designed two experiments. In the first experiment, the classification in the Corel data set is used, which was constructed manually by domain specialists. In the second experiment, the classification was made automatically using feature-based classification algorithms.

3.1 System Implementation

We integrated the proposed scheme into a prototype CBIR system, which adopts the query by example (QBE) mode. During the retrieval, the user only marks the retrieved images as relevant or irrelevant. There are six kinds of low level visual features used in the system: 256-dimensional color histogram in HSV space, first and second color moments in Lab space, 64-dimensional color coherence vector in LUV space, MRSAR texture feature, Tamura coarseness feature, and Tamura directionality. Their effectiveness has been proven in previous works (Liu et al. 2000, Lu et al. 2000). When

classification is used in retrieval, the semantic feature vector is defined by Equation (1).

3.2 Experimental Design

The image database we used consists of more than 10,000 images from the Corel dataset. It is a large and heterogeneous image set. Domain professionals classified all images in the dataset into semantic categories, each of which contains 100 images. Originally, there are 101 categories in the database. We merged some categories that share the same semantic meaning, such that each merged category depicts a distinct semantic topic. For example, category *Castles* in Disk07 and category *Castle II* in Disk08 are merged into one category, *Castle*. After that, we got a total of 79 categories.

We designed an automatic feedback scheme in the experiments. First, some query images are selected, and the ground truth is determined for each query image. After that, a reproducible relevance feedback for every query can be simulated based on the ground truth and the retrieval results of the system. At each iteration of the feedback, the system scans the first 30 retrieved images, and labels 3 irrelevant examples and at most 3 relevant examples, not including previous labelled images. Then the system uses these labelled examples and that from previous iterations to conduct relevance feedback.

Precision-scope curve is used to evaluate the image retrieval performance. Scope specifies the number of images returned to the user. Precision is defined as the percentage of retrieved images that are actually relevant. Results with and without classification information are referred to as "Class" and "Orig" respectively.

3.3 Result Using Manual Classification

In this experiment, all 79 categories in the database are used as classification information. Thus, a 79-dimensional vector is formed for each image as the semantic feature.

We selected 200 image examples as the query set. Most images are from the database, but 4 images are not. Then we collected judgments from 7 subjects as the ground truth. For every query example, each subject was required to scan the whole image set, and label all relevant ones based on his/her subjective judgment. In fact, the subjects selected different and varying numbers of relevant images for each query image. Therefore, the retrieval precision was calculated for each subject separately and averaged finally, as if there were 1,400 different queries.

The result of this experiment is presented in Figure 1, where "rf" stands for relevance feedback, the digit before it stands for the number of iterations in feedback. The semantic weight is set to 0.2 in case that classification is used. The result shows that the proposed method can improve the performance in this experiment. P(20) of "Orig" without relevance feedback is 0.16. After the second iteration of relevance feedback, P(20) of "Orig" is 0.27, while that of "Class" is 0.54.

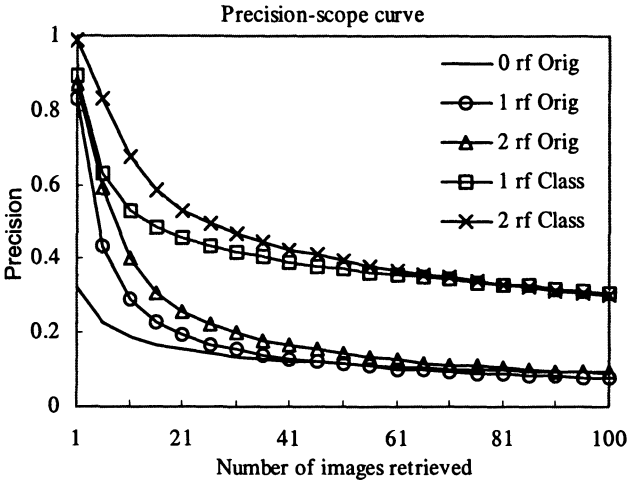

Figure 1. Average precision for the first two iterations.

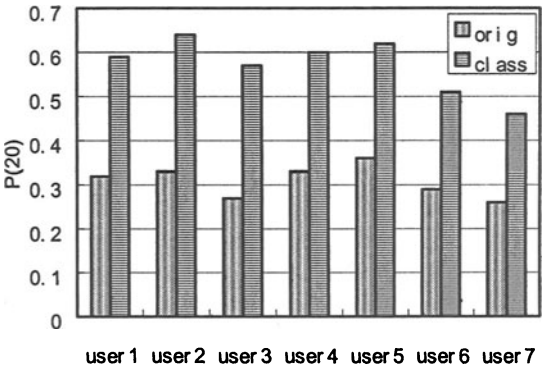

Figure 2. P(20) of different users.

Figure 2 illustrates the improvement of P(20) for different users. We can see that our approach makes obvious improvement for all 7 users. After investigating the data in detail, we also got some valuable observation. If most of the ground truth images of a query distribute in 1 or 2 categories, our method tends to make a greater improvement in precision. On the contrary, if the distribution is dispersive, our method tends to make a smaller improvement, or even make a little drop. In other words, if there is a good agreement between the user's intention and the image classification, our method will likely make improvement. Because the classification in Corel dataset was performed by domain professionals, it might match the perception of a common user very well. So much improvement is observed in the experiment.

3.4 Relationship between the Improvement and the Nature of Classification

Due to the subjectivity of human visual perception, certain classification can not satisfy all the different users and uses of images. This is why, in the above experiment, some queries of the total 1,400 queries are not consistent with the classification, although it was performed by domain professionals. To figure out how the nature of classification affects the efficiency of our approach. We did some quantitative analysis of the relationship between the performance improvement and the agreement between the classification and the user's perception. The improvement of the performance is measured in terms of the difference of P(20) after 2 iterations of relevance feedback, based on the method using classification and the original method only using low-level features. The classification and user's intention can be regard as two partitions of the image set. For a given classification, the partition is $\Theta_C = \{\theta_{C_1}, \ldots, \theta_{C_M}\}$, where M is the number of classes. The user's intention can also be considered as a classification, which classifies images into two classes, either relevant or irrelevant. And the corresponding partition is $\Theta_U = \{\theta_R, \theta_I\}$. The images in the same subset of a partition are regarded to be similar, the images from different subset dissimilar. The agreement between a classification and a user's intention is measured based on their agreement on the similarity or dissimilarity of image pairs. Inspired by Squire and Pun (1998), the agreement measure is defined as:

$$S(\Theta_C, \Theta_U) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X_{ij}(\Theta_C, \Theta_U) \tag{7}$$

where N is the number of relevant images in the database, i.e., only the relevant images are used in computing the agreement. This is because the similarity between the irrelevant images is hard to judge. $X_{ij}(\Theta_C, \Theta_U)$ is a binary variable which is 1 when relevant images i and j are in the same class of Θ_C. The value of S is between 0 and 1.

The agreement measure of all queries is quantified into 10 degrees, and the precision improvement is averaged for the queries belonging to the same degree. The precision is that of top 20 images at the 2nd iteration of relevance feedback. The analysis is based on the results of the experiment described in section 3.3. The relationship between the improvement and the agreement is illustrated as following.

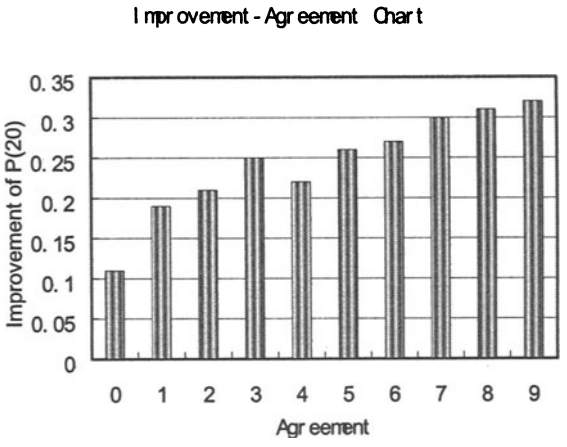

Figure 3. The relationship between improvement and agreement

From figure 3, we can find that more agreement between the classification and the user's perception, more improvement on the precision of the corresponding queries. Note that the agreement is based on statistics. The user's perception does not necessarily have explicit relation with the classification. For example, there is no explicit relation between the beach images and the categories of Asia, Europe, America, Africa, and Oceania. However, if most beach images are distributed in one or two categories, our approach will take advantage of this information and improve image retrieval accuracy.

3.5 Result Using Automatic Classification

In this experiment, we tested our method when using automatic image classification. Two SVM classifiers were trained to classify the images into semantic classes. Images were first classified to indoor or outdoor and outdoor images were further classified to man-made or natural. This method achieved an accuracy of 91.6% for indoor vs. outdoor classification, while 93% for man-made vs. natural classification. It is comparable to other classification algorithms (Szummer & Picard 1998, Vailaya et al. 1998).

In this experiment, we chose 30 out of 79 categories, totally 4,400 images. The criterion of selecting these 30 categories is mainly to guarantee the accuracy of image classification, and exclude those images unrelated to the classification, such as images of "flag", or "marble".

The classification made by domain professionals was not used in composing the semantic feature, but as the ground truth. That is, a retrieved image was considered a match if it belongs to the same category of the query image. The automatic classification was used to form a 4-dimensional semantic feature vector for each image. The semantic weight is set to 0.1. We tested 220 sample images chosen randomly from these 30 categories. To evaluate the performance, we averaged the precision over all queries. P(20) of "Orig" without relevance feedback is 0.28. After 2 iteration of relevance feedback, P(20) of "Orig" is 0.48, while P(20) of "Class" is 0.52. The improvement is small.

After analyzing the data, we found that the observation in the previous experiment still holds. Since the automatic classification is very coarse, too many images are located in one or two classes, and many queries from these classes can get less benefit from our scheme, and the performance also suffer from misclassification and inconsistency between some users' intention and the automatic classification.

4. DISCUSSIONS AND CONCLUSIONS

In this paper, we have proposed a flexible scheme to effectively combine the semantic classification with content-based image retrieval. Our approach automatically detects the correlation between the user's intention and the classification and takes advantage of it. Although any kind of semantic classification can be used, the advantage brought by our method does depend on the nature of the classification. If the classification conforms with the consensus of the users' intention well, an apparent improvement of the overall performance may be achieved. Experimental results on a large

heterogeneous image collection with manual classification show that the proposed method is effective.

We may further refine this method in the future. In Section 2, a fixed weight is used to combine the semantic feature and low-level features. Dynamic adjustment of the weight may make the approach more flexible. Furthermore, we may take account of the relationship between classes in image retrieval. In practice, this method may be used together with the traditional methods that use image classes as a filter.

ACKNOWLEDGMENTS

We thank Zhong Su, Fang Qian and Xiaofei He for constructive feedback.

REFERENCES

Bradshaw B. (2000) Semantic Based Image Retrieval: A Probabilistic Approach. Proc. ACM Multimedia, Los Angeles, California.

Baeza-Yates R. and Ribeiro B., editors. (1999) Modern Information Retrieval. Addison Wesley.

La Cascia M., Sethi S., and Sclaroff S. (1998) Combining textual and visual cues for content-based image retrieval on the World Wide Web. In Proc. IEEE Workshop on Content-based Access of Image and Video Libraries, Santa Barbara, CA.

Cox I. J., Miller M. L., Omohundro S., and Yianilos P. (1996) PicHunter: Bayesian Relevance Feedback for Image Retrieval. In Int. Conf. on Pattern Recognition, Vienna, Austria.

Cox I. J., Ghosn J., Miller M. L., Papathomas T. V., and Yianilos P. N. (1997) Hidden annotation in content based image retrieval. In Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries, pages 76-81.

Flickner M., Sawhney H., Niblack W., Ashley J., Huang Q., and Dom B. et al, (1995) Query by Image and Video Content: The QBIC System. IEEE Computer, 28(9).

Liu W., Sun Y., Zhang H.J. (2000) MiAlbum—A System for Home Photo Management Using the Semi-Automatic Image Annotation Approach. Proc. ACM MULTIMEDIA 2000, Los Angeles, California.

Lu Y., Hu Ch., Zhu X., Zhang H.J., Yang Q. (2000) A Unified Semantics and Feature Based Image Retrieval Technique Using Relevance Feedback. Proc. ACM MULTIMEDIA 2000, Los Angeles, California.

Pentland A., Picard R., and Sclaroff S. (1994) Photobook: Tools for Content-based Manipulation of Image Databases. In Proc. of the SPIE Conference on Storage and Retrieval of Image and Video Databases II, pages 34--47.

Rui Y., Huang T. S., Ortega M., Mehrotra S. (1998) Relevance feedback: A power tool in interactive content-based image retrieval. IEEE Trans. on Circuits and Systems for Video Technology, 8(5).

Rui Y. and Huang T. S. (2000) Optimizing Learning In Image Retrieval. Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), Hilton Head, SC.

Smith J. R. and Chang S. –F. (1996) Visualseek: a fully automated content-based image query system. In Proc. of ACM Multimedia 96, pages 87--98, Boston MA USA.

Smith J. R., Chang S. –F. (1997) Visually Searching the Web for Content. IEEE Multimedia Magazine, Vol. 4 No. 3, pp. 12-20.

Squire D. M. and Pun T. (1998) Assessing agreement between human and machine clusterings of image databases. Pattern Recognition, Vol. 31, No. 12.

Szummer M. and Picard R. W. (1998) Indoor-Outdoor Image Classification. IEEE Int. Workshop on Content-based Access of Image and Video Databases.

Wang J. Z. (2000) SIMPLIcity: A region-based image retrieval system for picture libraries and biomedical image databases. *Proc. ACM Multimedia, Los Angeles, California.*

Vailaya A., Jain A., and Zhang H.J. (1998) On Image Classification: City Image vs. Landscapes. *Pattern Recognition, Vol. 31, No. 12, pp. 1921-1936.*

BIOGRAPHIES

Hong Wu received his B.S. degree in computer science in 1993 from University of Science and Technology of China, and is now a Ph.D. candidate in National Lab. of Pattern Recognition at Institute of Automation, Chinese Academy of Sciences. His research interests include image processing, machine learning, and information retrieval, with a focus on content-based image retrieval.

Mingjing Li received his B.S. in electrical engineering from University of Science and Technology of China in 1989, and Ph.D. in Pattern Recognition from Institute of Automation, Chinese Academy of Sciences in 1995. He joined Microsoft Research Asia in July 1999. His research interests include handwriting recognition, statistical language modeling, search engine, and multimedia information retrieval.

Hong-Jiang Zhang received his B.S. from Zhengzhou University, China in 1982, and Ph.D. from the Technical University of Denmark in 1991, both in electrical engineering. His research interests include video and image analysis and processing, content-based image/video/audio retrieval, media compression and streaming, computer vision and their applications in consumer and enterprise markets. He has published over 120 articles in the above area. He is a senior member of IEEE, also serves on the editorial boards of 5 professional journals and a dozen committees of various international conferences.

Wei-Ying Ma received the B.S. degree in electrical engineering from the national Tsing-Hua University in Taiwan, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California at Santa Barbara. From June 1997 to March 2001 he was a researcher in the Hewlett-Packard Laboratories at Palo Alto. He is currently a research manager of Media Management Group in Microsoft Research Asia. He is an associate editor for the Journal of Multimedia Tools and Applications, and has served on organizing and program committees of many international conferences.