Survey on Image Content Analysis, Indexing, and Retrieval Techniques and Status Report of MPEG-7

Zijun Yang and C.-C. Jay Kuo

Integrated Media Systems Center
Department of Electrical Engineering-Systems
University of Southern California, Los Angeles, CA 90089-2564
Email:fzijun,cckuog@sipi.usc.edu

Abstract

Multimedia database management has been intensively studied recently due to the rapid growth of multimedia data and the demand of their access over the Internet. The new member of the MPEG (Moving Picture Expert Group) family, called the "Multimedia Content Description Interface" or MPEG-7 in short, will extend the limited capabilities of proprietary solutions in identifying multimedia contents that exist today. State-of-the-art technologies and systems will be evaluated and merged to specify a standard to describe multimedia contents in such a process. In this review paper, a comprehensive survey of image database management techniques and systems is performed, and the status in the MPEG-7 standardization process is reported. Based on this study, future trends and promising research directions are also predicted.

Key words: Image content analysis, Feature extraction, Image retrieval, MPEG-7, Descriptors, Description scheme, Description definition language

1. Introduction

Advances in modern multimedia technologies have led to huge and ever growing archives of images, audio and video in diverse application areas such as medicine, remote-sensing, entertainment, education and on-line information services. While audio-visual information used to be consumed by human beings, there is an increasing number of cases where the multimedia information is created, exchanged, retrieved and re-used by computational system [58]. This is similar to what occurred in the early computer development stage, during which the huge amount of alpha-numeric data in- creased rapidly and many practical issues in the DataBase Management System (DBMS) arose [85]. In the past, DBMS was designed to organize alpha-numeric data into structured records which are indexed by key attributes so that information retrieval and storage could be done conveniently and efficiently. However, traditional DBMS does not work well for multimedia data due to difficulties in several aspects, which include the diversity of the data type (e.g. image, video, audio), the large capacity of the unit record, and the lack of semantic meaning of the data at the physical level. To exploit the full benefit of the explosive growth of multimedia data, there is a strong and urgent demand in developing efficient techniques for their storage, browsing, indexing and retrieval.

The most natural way is manual annotation, which most early work focused on. However, there exist two major disadvantages, especially when the size of image collections grows large. One is the vast amount of labor required in manual image annotation. The other difficulty, which is more essential, results from rich contents in images and the subjectivity of human perception. Content-Based Image Retrieval (CBIR), which extracts visual features such as color, texture and shape from images automatically, is dedicated to overcome these difficulties. There are three fundamental blocks in a CBIR system, i.e., visual feature extraction (descriptors), image management system design (description scheme), a language to specify descriptors and description

schemes (description definition language). Fig. 1 shows an image content management system architecture consisting of these three basic components. The structure can be extended to general multimedia content description systems.

ISO (International Standard Organization) /MPEG (Moving Picture Experts Group) [58], [59] has started a new work item to provide a solution to describe multimedia data contents and support content-based multimedia management. The new member of the MPEG family, called “Multimedia Content Description Interface” or MPEG-7 in short, will extend limited capabilities of today's proprietary solutions in identifying media contents, notably by including more diverse data types. In other words, MPEG-7 will specify a standard set of descriptors to describe various types of multimedia information. The scope of MPEG-7 is illustrated in Fig. 2, which shows a highly abstract block diagram of a possible MPEG-7 processing chain [58]. This chain includes feature extraction (analysis), the description itself, and the search engine (application). MPEG-7 does not specify detailed algorithms of feature extraction and implementations of certain applications but the description which provides a common interface between these two stages. A hypothetical MPEG-7 chain in practice is shown in Fig. 3. Possible practice covers various cases of professional and consumer applications [57] such as education, journalism, tourist information, entertainment, investigation services, geographical information systems,etc.

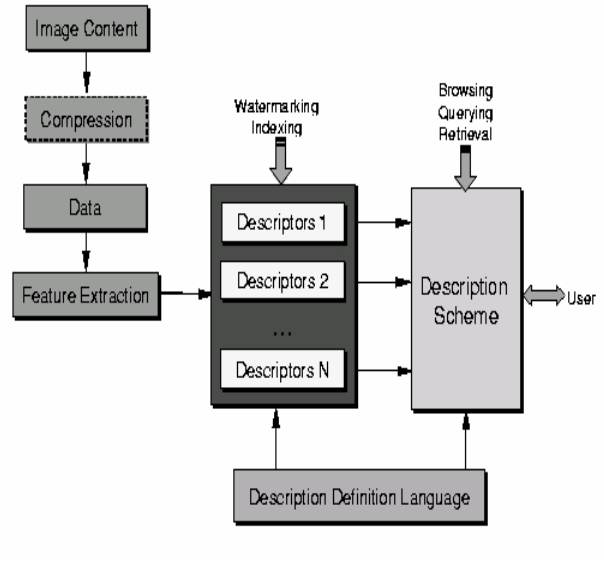

Figure 1. The CBIR system architecture with three Component

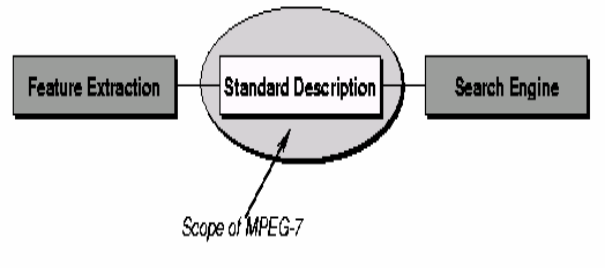

Figure 2. Scope of MPEG-7

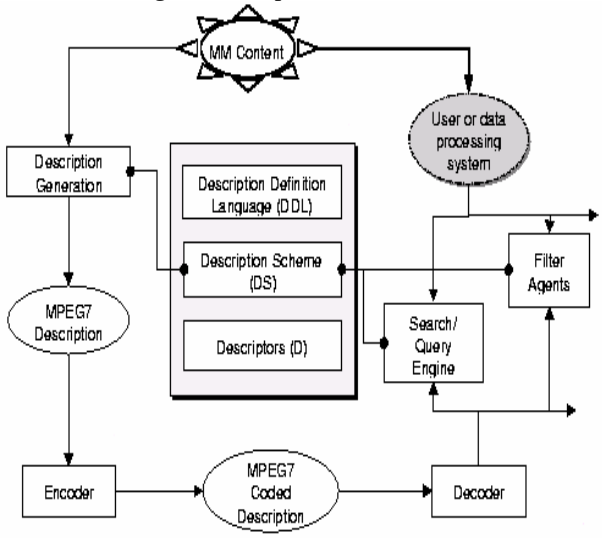

Figure 3. An abstract representation of possible application using MPEG-7

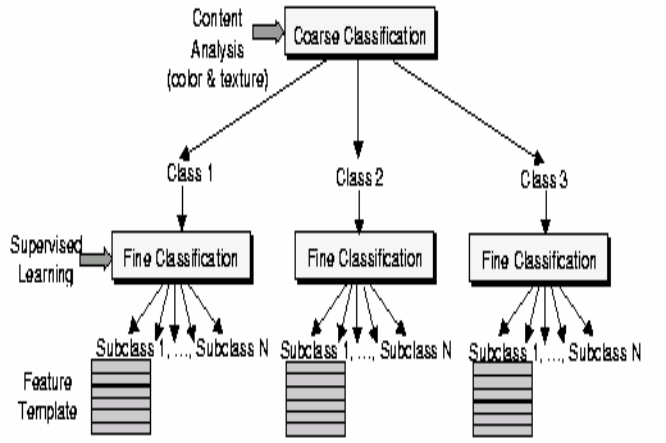

Figure 4. The hierarchical classification structure based on content analysis supervised learning

The characteristics of multimedia data are often represented by features such as the color of an image, the pitch of an audio clip, the meaning of a speech segment, etc. A descriptor is a representation of a feature. One feature can be associated with multiple descriptors. For example, the average color, the dominant color, and the color histogram can all be descriptors of the color feature. MPEG-7 will also standardize ways to define other descriptors as well as structures (Description Scheme) of descriptors and their relationships. A

description scheme should support (a) specific searches, where the query is well formulated with appropriate constraints, (b) browsing to quickly understand contents of a database or one of its subsets and (c) navigation, such as a hypermedia paradigm which allows users to traverse the image space using links, and so on. In other words, description schemes refer to subsystems which solve image classification and organization problems, describe image contents with specific indexing structures, or retrieve relevant images from an image database based on user's interest. The combination of descriptors and description schemes is called “description”, which shall be associated with the content itself to allow fast and efficient searching for material of user's interest. MPEG-7 will also standardize a language to specify the description scheme called the Description Definition Language (DDL). DDL should allow modification, extension and identification of descriptors and description schemes. Techniques such as image content analysis, effective feature indexing, and interactive retrieval are critical to the design of a multimedia database management system.

Since an extensive survey of current content-based image retrieval paradigms already has been made by Rui, Huang and Chang in [81], we will focus primarily on image content analysis and recent developments on feature extraction, indexing and retrieval techniques in this paper. In particular, we will report current MPEG-7 activities and review work which has been proposed and/or adopted in the eXperimental Model (XM) or the Core Experiments (CE) of MPEG-7 [74].

This paper is organized as follows. Various visual feature extraction algorithms are examined in Section 2. Their corresponding representation and matching techniques are also discussed. To facilitate fast search in large-scale image collections, issues of effective and efficient image retrieval system design are addressed in Section 3, where various techniques, including image analysis, classification, indexing and retrieval methods, are evaluated. The state-of-the-art technologies and systems adopted in XM or CE of MPEG-7, together with their characteristics are described in Section 4. The representative description definition languages currently considered by MPEG-7 are presented in Section 5. Future trends and promising future research directions are suggested in Section 6. Finally, concluding remarks are given in Section 7.

2. Visual Feature Extraction: Descriptors

Since a tremendous amount of effort is demanded by manual indexing, automatic indexing based on visual features has been widely studied as an attractive alternative. Image features are distinguishing primitive characteristics (or attributes) of an image, which serve as the core building block for modern CBIR systems. It is observed that low level visual features such as color, texture and shape are sometimes, but not always, correlated well to image semantics. Other features such as visual objects, sketch, and spatial relationship are also considered in MPEG-7 requirements. Due to image content varieties and diverse application subjectivity, there exists neither a universal feature for all images nor a single best representation for a given feature. The “best” features may change from images to images and from applications to applications.

The set of descriptors should allow a great flexibility to meet general, functional, visual specific and coding requirements [59]. General requirements include the type of features, abstraction levels for multimedia material, cross-modality, multiple descriptions, descriptor priorities, hierarchy, scalability and so on. Functional requirements include retrieval effectiveness, efficiency, similarity- based retrieval, associated information, intellectual property information, etc. In visual specific requirements, the type of features, data visualization using the description, visual data formats and visual data classes are considered. Coding requirements include description efficient representation, description extraction and robustness to information errors and loss.

The color feature is one of the most widely used visual features in image retrieval. It is relatively robust to background complication and independent of image size and orientation. Descriptors for the color feature are mostly statistics of color distribution, e.g., the color histogram, the average color and color moments. The selection of color space and color quantization schemes in calculating these statistical value can greatly influence the efficiency of the underlying descriptors. Texture is an important attribute of image for surface and object identification. It has been used to classify and recognize objects and scenes. Shape is important in its own right for objects, such as object detection, representation and motion. As a matter of fact, in most cases human beings pay more attention to some specific interesting objects or areas instead of the whole pictorial scenes.

2.1 Color Descriptors

Among various feature types of color descriptors, the dominant color means the most prominent color representation and has been considered one of major descriptors in MPEG-7 XM because of its simplicity and association with human perception. Ohm and Makai proposed a simple dominant color descriptor [64] which is defined by the mean value of a color cluster. In other words, the color index of the quantizer cell nearest to the centroid of the color cluster is used to represent the dominant color. Since human vision perception is more sensitive to changes in smooth regions than in detailed regions, Manjunath et al. proposed a method called the Variable-Bin Color Histogram (VBCH) [23] which utilizes peer group filtering [22] and weighting to assign different emphases on different areas. Then, an agglomerative clustering algorithm is performed on cluster centroids to further merge close clusters and obtain the dominant color. A Similar Color Image (SCI) was studied by the IBM Almaden Research Center [38] to produce a single color descriptor and a matching approach desirable in applications insensitive to the color position. Instead of using a single dominant color, Mottaleb and Krishnamachar [1] took several dominant colors to represent an image. This approach retains eight dominant colors.

The color histogram is the most basic color content representation which describes statistical color distributions by quantizing the color space. It can be applied to any image shapes. Three quantization types, i.e. linear, nonlinear and lookup table quantizers, were proposed by Ohm and Makai [64]. If quantization is actually performed on a visual item, the number of pixels that fall into each quantizer cell can be determined by a histogram descriptor. The color histogram descriptor proposed by IBM [38] is computed over any area containing image pixels, which can be the whole image or an image region of any shape such as the rectangular subregion or the segmented object. The smoothing procedure is also conducted to produce better results in case of color dithering. Since a single global histogram of the whole image is not very effective, a multiple color histogram descriptor capturing the local spatial variation of colors was proposed by Mottalab and Krishnamachar [1]. Each image is divided into N partitions horizontally and M partitions vertically. Then, color histograms of N ┐M rectangular regions are obtained. The selection of the number of rectangular regions should be based on the dimensions (size) of underlying images. A generalized image histogram, accepting a wide variety of color models and compressed bit streams as well, was proposed by Won et al. [103], [102]. In this approach, a set of pixels or a single pixel can be adopted as a basic pixel-unit for the histogram computation. Then, a linear quantization for histogram generation is adopted to efficiently represent the image.

To extend visual color features from still images to video data, a histogram-based color descriptor for a Group Of Frames (GOF) was proposed by Tekalp et al. [32], [31]. GOF is defined as a set of frames that have been clustered according to a certain criterion. The intersection histogram and the median histogram descriptors are used to describe visual features of GOF. Histogram intersection defines a scalar measure for comparing two histograms, and yields the number of pixels that share the same color. The median histogram is a good representative of the average color distribution of a collection of frames. Another descriptor called the “super histogram” was proposed by Dimitrova et al. [25], [26] for video content analysis and classification. The method computes the color histogram for individual shots and then merges the resulting histograms into a single cumulative histogram called a family of histograms based on a comparison measure. A few families of histograms can be obtained to represent the entire TV program.

Pixel-based color histogram descriptors do not incorporate any spatial information and, therefore, can not be used to differentiate objects with different sizes or shapes. To solve this problem, Beek et al. [7] proposed a scalable image-blob histogram descriptor, which is histogram-based yet generalized to include spatial information. Instead of encoding the frequency distribution of attributes, the image-blob histogram descriptor encodes the relative size and distribution of groups of pixels with uniform color without the need of segmentation. The descriptor can be easily extended to describe texture features. Some parameterized color distribution descriptors were investigated to take into account spatial information [19], [40], [95]. Cieplinski proposed a color descriptor [19] that is applicable to images and objects extracted from images. The descriptor consists of the mean value of the object color and its covariance matrix. Jung et al. proposed a color distribution feature [40] in terms of subregions in a whole image. First, a video frame is divided into high and low entropy regions. Color descriptors

incorporating region information are further obtained. Tabatabai proposed a color feature representation [95] to describe color properties of visual objects. The visual object, either segmented automatically or semiutomatically, is used to denote a semantic object.

There are several commonly used color spaces, such as RGB, CIE, HSV, HSI and Munsell color spaces. The RGB color space is extensively used to represent images. Other spaces such as CIE, HSV, correlate better with human perception. Descriptors in different color spaces were considered and evaluated in [64], [38], [42]. A descriptor for quantized colors with the HMMD color model, which is claimed to be more uniform than the HSV color space, was proposed by Kim et al. [42]. The HMMD color model was developed from the RGB and the HSV color spaces. It consists of natural parameters such as hue (H), tint (Min), shade (Max) and tone (Diff = max – min). A color is represented by (hue, max, min)or (hue, diff, sum). Then, the HMMD space is uniformly quantized along with each parameter h, max, min, and diff or quantized by using two different methods according to achromatic and chromatic colors.

2.2 Texture Descriptors

Texture is a term which defies a rigorous, complete and formal definition. However, even without a clear definition, no one would argue that the Human Visual System (HVS) relies heavily on texture perception for image interpretation and analysis. The notion of texture appears to depend upon three ingredients [70]. First, some local "order" is repeated over a region which is large in comparison to the order's size. Second, the order consists in the nonrandom arrangement of elementary parts. Third, there are roughly uniform entities having approximately the same dimension every- where within the textured region. In other words, texture is generated by one or more basic local patterns that are repeated in a quasi-periodic manner over some region or visual images that possess some stochastic structure. Since texture provides important characteristics for surface and object identification, they have been extensively studied and applied in industrial monitoring of product quality, remote sensing of earth resources and medical diagnosis with computer tomography.

Descriptors for texture features can be classified into two categories: statistical model-based and transform-based. The first approach explores the gray level spatial dependence of textures and then extracts meaningful statistics as texture representation. Co-occurrence matrix representation proposed by Aksoy and Haralick [4] analyzes the variance of the gray level co-occurrence matrix to classify various texture collections. Line contents of images can be used to represent texture of the image. Aksoy and Haralick also studied the Line-Angle-Ratio statistics [4] by analyzing the spatial relationships of lines as well as the properties of their surroundings. Tao and Dickinson recognized different texture patterns by using a modied gradient indexing technique called the Local Activity Spectrum [97]. However, the above statistical approaches do not exploit the sensitivity of the human visual system to textures. Motivated by psychological studies in texture visualization, Tamura et al. [96] tried to determine relevant features used in texture perception. A human texture perception study conducted by Rao and Lohse [72] indicates that the three most important orthogonal dimensions are "repetitiveness", "directionality", and "granularity and complexity". Liu and Picard [50], Niblack et al. [62], [63] used contrast, coarseness and directionality models to achieve texture classification and recognition.

Statistical texture descriptors are not readily to be applied to texture modeling in the compressed domain. Wan and Kuo [100] extended texture feature extraction to the compressed domain by analyzing the energy distribution based on AC coefficients of the Discrete Cosine Transform (DCT), which is adopted in JPEG (Joint Picture Expert Group) still image compression standard. Other various transforms can be used in transform-based texture descriptors. The Fourier-Mellin transform was proposed to classify rotated and scaled textures by Alata et al. [5], where the combination of the parametric 2-D spectrum estimation method called HMHV (Harmonic Mean Horizontal Vertical) and the Fourier-Mellin transform was adopted. It is however well known that the Fourier- or DCT-based descriptors can not characterize textures of different scales effectively. The Gabor and/or wavelet transforms were proposed to overcome this difficulty. The Gabor filters proposed by Manjunath and Ma [52] offer texture descriptors with a set of "optimum joint bandwidth". The wavelet transform offers a set of frequency/space localization bases which can exactly reconstruct the original signal based on wavelet coefficients. A tree-structured wavelet transform presented by Chang and Kuo [18] provides a natural and effective way to describe textures which have dominant middle or high

frequency subbands.

Recent work attempts to integrate statistical analysis with the wavelet decomposition [86], [61]. A statistical characterization of texture images based on a model represented by an overcomplete complex wavelet frame was investigated by Simoncelli and Portilla [86]. The characterization consists of local autocorrelation of coefficients in each subband, local autocorrelation of coefficient magnitudes, and cross-correlation of coefficient magnitudes at all orientations and adjacent spatial scales. In [61], Nevel developed a method which relies on matching the first and the second-order statistics of wavelet subbands. It goes beyond simple marginal matching, attempting to match correlation coefficients of subbands of interest as well.

Since texture patterns are important features for object recognition and description, MPEG-7 is targeting to incorporate texture features with spatial information [20]. Two classes of textures, i.e. the spatial image intensity distribution of textures and homogeneous textures have been recommended for core experiments. The Ricoh company proposed a spatial edge distribution [76] and spatial texture distribution descriptors [77]. Spatial edge distribution is abstract information to describe outlines of objects while spatial texture distribution describes where and which texture exists. An image is first partioned into some image blocks. For each block, the amount of edges or the centroid of edge areas, is calculated for four directions, i.e. **0^{o} , 45^{o} , 90^{o} and 135^{o}** . Texture for each block is extracted by using the co-occurrence array, and thirteen features can be calculated from co-occurrence array elements. Mottaleb used the histogram of edge directions [56] as a basis for deriving descriptors, which is similar to spatial edge distribution descriptors. The two descriptors will be merged in later stages.

Many methods have been developed to describe texture patterns and researchers attempt to combine them to derive multiple descriptors. Manjunath et al. proposed a homogeneous texture descriptor [106] which has the Perceptual Browsing Component (PBC) and the Similarity Retrieval Component (SRC). PBC provides a high level characterization of textures and SRC is computed by convolving the image with a set of filters tuned to detect image features at different scales and orientations. Ohm and Bunjamin proposed a composite texture descriptor [65], which defines the meanings of the frequency decomposition structure, quantization of coefficients and representation of statistics. A texture descriptor based on the human visual system (HVS) was proposed by Ro et al. [78]. The texture is described in Radon space to fit HVS behavior. Furthermore, by using the matching pursuit [79] in the Radon space, a texture descriptor that compactly represent the texture feature can be obtained. The matching pursuit method gives the content feature. That is, a small number of atoms after decomposition represents the texture feature of an image.

2.3 Shape Descriptor

Several qualitative and quantitative techniques have been developed in characterizing the shape of objects within an image. Shape features are useful for classifying objects in a pattern recognition system and for symbolically describing objects in an image understanding system. Some of these techniques apply only to binary images while others can be extended to gray level images. On one hand, the shape is an important feature in object representation and recognition. On the other hand, it is a great challenge to extract a set of accurate yet simple shape representations to specify an object since the shape is a projected result of a 3D object onto a 2D plane. For video data, object shapes and motion are often combined in object representation and analysis. For still images, only object shape descriptors are relevant.

According to different applications and requirements, MPEG-7 clusters shape descriptors into two groups addressing different functionalities [24]. The first one addresses similarity-based retrieval for simple pre-segmented shapes, defined by a closed contour. The requirement here is that the solution should be scale- and rotation-invariant and should be robust to small non-rigid deformations, for example, due to non-rigid motion. The emphasis here is on perceptual similarity. The second one addresses similarity-based retrieval for complex shapes, defined as a sum of disjoint binary regions. Good examples of shapes from this class are trademarks, or the character set. The requirement here is that the techniques should be scale- and rotation-invariant, but the requirement of a non-rigid deformation is not imposed.

To begin with, let us review several proposals in the first group. Bober [9] proposed a curvature scale-space representation of the shape, the outline and the curve. The descriptor is applicable to segmented objects in the image, a part of objects, curves, etc. The shape is represented by a list of 2D

vectors, each vector corresponding to a zero-crossings of its position on the run-length of the curve. Muller and Ohm [60] proposed a wavelet-based contour descriptor. A reduction of initial database contours is carried out to reduce the complexity by using the modification ratio of a contour dened by the ratio of the inscribed diagonal to the maximal diagonal. The ratio describes whether a contour is thin (needle like) or circular and is quantized into 16 discrete steps. A orientation-invariant descriptor was proposed by Kim and Kim [44] by using a set of Zernike moments. The reason for choosing orthogonal Zernike moments is that they possess a useful rotation-invariant features [41].

Besides offering rotation-invariant properties, the shape descriptor proposed by Kim and Kim [43] offers a multi-level contour representation capability. A shape can be represented by two eigenvectors from the covariance matrix of spatial position data [34]. These two eigenvectors point in the directions of the maximal and the minimal region spreads which are very effective characteristics of the shape.

The contour of extracted objects may vary with the orientation, zooming of the camera and the spatial position of the visual object within the picture. A generic contour representation of the shape is not a suitable format for matching in such situations. A normalized contour representation was proposed by Tektronix Inc. [98] to overcome the difficulty. This descriptor is accurate in describing simple shapes with a large number of corners. Such a descriptor can tackle not only rotation but also small non-rigid deformation.

Some shape descriptors can be extended to describe complex object contours. It is worthwhile to point out that the descriptor using the Zernike moment [44] can be also used to describe the complex geometric shape of a device-type trademark [45] (i.e. a mark that contains graphical or figurative elements only). The Multi-Layer Eigen Vector (MLEV) concept proposed by Kim and Kim [43] can be enhanced for complex shape description by calculating transformation invariant features with multi-level eigenvectors obtained by sub-dividing regions repetitively.

All proposals mentioned focus on descriptors for segmented objects. However, in many cases, it is extremely difficult to extract objects from images automatically. Mahmood [51] proposed an approach to achieve similarity retrieval in non-segmented images as well as complex shape description. In this proposal, a Location Hashing Tree (LHT) is used to organize the feature information from images and index the image database accordingly. Location hashing is based on the principle of geometric hashing. It determines relevant images in the database and regions within them that are most likely to contain a 2D pattern query without incurring detailed search. The geometric hashing technique was introduced for the model indexing problem in object recognition by Lamdan and Wolfson [107]. It has so far been applied to recognize the object depicted in an isolated region in an image without a systematic search of a library of model objects.

3. System Tools: Image Classification and Retrieval Techniques

Visual feature extraction applied to content-based image retrieval has been thoroughly studied for the last 5 years. Most work concentrates on low level visual features such as color, shape, texture, etc. and adopts a feature-based image retrieval approach. Examples include the QBIC system [62], the RetrievalWare system [29], the Virage system [6], the VisualSEEK and WebSEEK system [88], the Blobworld system [13], the Photobook system [54], the Mars system [80] and the USC system [111]. The application of these systems to real world problems is however limited due to the ignorance of content varieties and the lack of semantic meanings in extracted features. Specific low-level image features may provide a solution to image retrieval in some applications (e.g. with respect to a pre-selected image database), but may have a problem in handling other applications. Generally speaking, there exists neither a universal feature for all images nor a single best representation for a given feature. Even with a modification in the distance calculation [54, 80, 111], retrieval results are not always satisfactory.

To improve the image retrieval performance, it becomes clear that image classification tools [10, 11, 14, 21, 36, 39, 49, 53, 73, 82, 94, 99, 101, 110, 112] should be used in combination with the feature-based retrieval technique. That is, it is desirable to classify images in the image database into several categories via image classification first. Then, we will decide which features and descriptors to be used in the retrieval phase. This topic will be discussed in Section 3.1. Another important research area in image retrieval is the use of user's feedback to improve search results since the one-shot query often fails in practice. Research work on interactive image retrieval will be reviewed in Section 3.2.

3.1 Image Classification

Image classification helps the selection of proper features and descriptors for the indexing and retrieval purpose. It enhances not only the retrieval accuracy but also the retrieval speed, since a large image database can be organized according to the classification rule and search can be performed within relevant classes. Current work of image classification relies on either low-level features or heuristical rules. A feature-space organization technique that provides flexibility to support different access methods while being extensible to large image databases is highly in demand.

One common way to achieve classification is via clustering. Clustering has been extensively used in text-based information management systems for retrieval as well as classification [39]. A summary of document clustering techniques used to improve collection search was given by Salton and Araya in [82]. In the work of Cutting et al. [21], a technique called scatter/gather was proposed for dynamically clustering document collections to facilitate the browsing of large text-based information sources. Clustering was also utilized in hierarchical network search engines such as HyPursuit [101], where hypertext documents are clustered at different sites to structure the information space for browsing and searching. Several hierarchical network search engines address the issues of scalability in terms of storage and network communication requirements for the task of resource discovery. For details, we refer to the HyPursuit system by Weiss et al. [101] and the Harvest system by Bowman et al. [11].

There are numerous image classification tools proposed to classify images for a particular application. For example, the problems of “texture classification” and “human face classification” have been studied for years. Different image features are extracted for differing applications. Examples of image classifiers include the k-nearest neighbor, the decision tree, the Bayesian net, the maximum likelihood analysis, the linear discriminant analysis, the neural network, etc. However, relatively less work has been done on the classification of general images and the organization of distinctive image features.

To make the classification simpler, researchers have studied to cluster images into indoor and outdoor picture groups. Yu and Wolf [68] presented a one-dimensional Hidden Markov Model (HMM) for indoor/outdoor scene classification by learning a statistical template from examples. The HMM along vertical or horizontal segments of specific scene layouts, such as sky-mountain-river-scenes, is trained. A knowledge-based scene classification method was adopted by Lipson et al. [49] in constructing a congurable recognition system. Instead of building a specific scene class detector, a model template that encodes the common global scene configuration structure by using qualitative measurements is hand-crafted for each category. Szummer and Picard [94] exploited multiple feature combination to improve the indoor-outdoor image classification performance by 10-15%. Three types of features: one each for color, texture and frequency information, are integrated to compute features on subblocks, classify these subblocks and then combine these results in a way reminiscent of “stacking”.

To avoid the drawback of manual creation of templates, a learning scheme that automatically constructs scene templates from a few examples was studied by Ratan and Grimson [73]. The learning scheme was tested on two scene classes with promising results. Carson et al. [14] proposed a new representation for images. Each image was thought to consist of several blobs where each blob is coherent in its color and texture components. All blobs in the training data of 14 image categories were clustered into about 180 “canonical” blobs by using a multivariate Gaussian model of a diagonal covariance matrix. Vellaikal and Kuo [99] proposed a hierarchical clustering technique for natural image database organization and summarization, where images were grouped via low level feature clustering. Different types of hierarchical agglomerative clustering techniques were studied to organize features for image group summarization. Huang et al. [36] examined an automatic hierarchical image classification scheme. They selected an initial set of low-level features, and subsequently performed feature-space reconfiguration by using the singular value decomposition to reduce noise and dimensionality. A mixture decomposition technique was described by Medasani and Krishnapuram [53], which automatically found a compact representation of images in terms of categories and applied it to the problem of database organization for efficient retrieval. This algorithm can determine the appropriate number of categories required to model an image database automatically. The least trimmed square approach was used to limit the influence of outliers and noise on class/component parameters.

All above classification work is based on low level feature clustering. It is well known that there is a big gap between low level features and semantic meanings. Some researchers studied the knowledge-based approach to organize image collections. Bouet and Djeraba [10] proposed an image organization scheme in a visual retrieval system by using the "concept-based" query. The first extension beyond textual and visual feature combinations in this system was to enable users to express their queries with more semantics and to store their concepts in the database. The second extension introduced was the use of the knowledge discovery approach to learn concepts in queries. The extension permitted the system to classify automatically a new image during its insertion in the large image collection and obtain results with more semantics so that the overall retrieval result can be improved.

The use of image content analysis for understanding image content, organizing image database and choosing appropriate low level features and semantic meanings in the image indexing/retrieval application was integrated into a single framework by Yang and Kuo [110], [108]. A hierarchical classification procedure was proposed by using two-level (coarse- and fine-level) classifications. The coarse-level classification is based on semantic meanings while the fine-level classification is based on low-level image features. In coarse classification, color and edge information analysis was utilized to summarize image collections with pre-selected image models. In fine classification, a supervised training algorithm based on multiple feature templates was adopted to refine the classification result furthermore within each coarse class [109]. This concept is illustrated in Fig. 4.

3.2 Interactive Image Retrieval

The image retrieval engine usually consists of a human-user interface, an image analysis unit and a matching mechanism. The search of desired images within a large collection of images can be based on text and/or image query. The image-based query (or query-by-example) is also known as content-based image retrieval. In most CBIR systems [6, 13, 29, 62, 88], users are allowed to select one or multiple query images, features of interest, the distance measure and weighting of each feature to perform the matching in the interface module. Then, a nearest-neighbor type of algorithm is performed in the matching module. The image analysis part in existing CBIR systems is very simple, i.e., to extract features specified by users. The major shortcoming in these systems is that there exists a big gap between high level semantic concepts and low level features. It is extremely difficult to describe high level semantic concepts with image features only. Interactive retrieval based on user's feedback provide a promising solution to this problem. In some systems [105], [8], [35], [80], [111], interactive learning is performed at the retrieval stage to refine the query and the metric at run time, which we call "short term learning". The improvement of image database organization has been conducted in [92], [54], [110] by refining the image groupings and classification, which is known as "long term learning".

Wood et al. [105] exploited user's knowledge to a higher extent by employing relevance feedback to iteratively refine queries at run time and achieve a short term learning goal. Subjects of interest were chosen via selection of regions from pre-processed and segmented images, thus allowing access to object-specific local information which is not possible with a global pattern-matching approach. After an initial retrieval attempt, feedback is given in the form of acceptance or rejection to each retrieved image. Bhanu et al. [8] proposed a retrieval system that continuously learned the weights of features and selected an appropriate similarity metric based on user's feedback given as positive or negative image examples. The method enables the image retrieval engine to learn the feature relevance, which is adaptive to the dynamic query process. An interactive retrieval system that combined supervised learning with color correlograms was proposed by Huang et al. [35]. In this system, both learning the query and learning the metric were used to improve the system performance. The Mars system developed by Rui et al. [80] supported interactive retrieval with user's relevance feedback, where a multimedia object model was represented by raw data of the object, the set of features associated with the object, the set of representations for a given feature, the set of similarity measures and the set of realizations of similarity measures. The top layer relevance feedback selected the best combination of the similarity measure and the feature set by user's ranking of retrieval results. The bottom layer feedback improved the system performance by using user's feedback in the realization of the similarity measure layer. A query system integrating multiple query seeds and relevance feedback was proposed bt Yang et al. [111]. In this

work, an interactive process was used to determine the meaning of “similarity” gradually with users in the loop. Users imposed a query with an image set, where multiple features (e.g. color, texture, shape) were adopted in the similarity computation with an equal weighting initially. A set of images could be retrieved accordingly. Then, retrieved images were classified as relevant or irrelevant by users and the weights of features were adjusted adaptively based on the relevance feedback.

The above short term learning systems achieve only user-oriented retrieval tasks, but do not contribute to database system design. Long term learning systems improve the database system structure by updating the image organization and classication upon user's evaluation. Squire [92] employed human partitionings of an image set to improve the organization of an image database. This system updated the similarity distance measure with user's interactivity. Minka and Picard [54] proposed an interactive learning approach by using a `society of models'. Their work utilized many data-dependent, user-dependent, and task-dependent features in a semi-automated system. A learning algorithm was presented for selecting and combining groups of data, where groupings can be induced by highly specialized and context-dependent features. Features are further selected according to user's positive and negative answers to groupings for a query. An adaptive image organization system was proposed by Yang and Kuo [110], [109]. In this framework, hierarchical classification is continuously updated by learning users' feedback and interactivity in terms of a long period.

4. Description Schemes

In MPEG-7, a Description Scheme (DS) [58] is used to specify structures and semantics of the relationship between several components, which may be Descriptors (D) or DS. The description scheme shall be associated with the image content to allow fast and efficient search of images of user's interest. The design of description schemes should satisfy general, functional, visual specific and coding requirements [59] as we described in Section 2. High level description schemes that utilize some type of tree- or graph-structure and have the ability to describe relationships between individual descriptors are favored in MPEG-7. Some description scheme proposals submitted to MPEG-7 are reviewed in this section.

ACTS-DICEMAN [2] proposed a scheme similar to the table of content and the table of indices used in books. Two hierarchical structures were represented by two trees, i.e. the region tree and the object tree. The region tree describes the spatial organization of a given image, where each node represents a connected component in the space domain called a region. The object tree is composed of a list of objects judged as being of potential interest during the indexing process. This tree is composed of objects with semantic meaning. An example of references from the object tree to the region tree is given in Fig. 5. An object can be jointly represented by several different regions. The corresponding descriptors typically define the type, identity and/or activity of the object. Moreover, for the indexing purpose, the most important functionality of the object tree is to relate its objects to regions of the region tree. Note that the region tree is a signal-based representation of an image while the object tree provides a semantic description. These two types of descriptions are useful in answering queries but may not be efficient in supporting a simple browsing function where the user wants to have a quick overview of the image collection.

Cha et al. [15] proposed a scheme based on domain ontology to represent the situational meaning of an image. By situation, we mean something related to questions in one's mind such as “What is the image about?”, “Where is the location of an object in an image?”, “What are people in the image doing?” or “Who is the one standing on the platform?”, etc. The description scheme is able to answer this type of questions effectively. Typical examples include: “I want images of Bill Clinton at the White House” and “Give me images of Michael Jordan dunking in games”. Cha gave a definition of ontology as the specification of relationships among notions or concepts of words through hierarchical classification. The ontology [69] can describe concepts in a domain of discourse by several abstraction levels of description and well-defined and unambiguous
semantics. The proposed solution represents features related to situational meaning of a still image, and consists of five descriptors which are geographical, component (different objects in an image), context, relational (spatial relationship) and temporal information.

In the joint proposal of Columbia University, IBM and AT&T [68], an object-based image DS represented by eXtensible Markup Language (XML) [30] was presented. The proposed DS consists of several basic components: objects, the object hierarchy, the entity relation graph and the

feature structure. Each image description includes a set of objects. Objects can be organized in one or more object hierarchies. The relationships among objects are expressed in one or more entity relation graphs. Such an image DS structure is illustrated in Fig. 6. Each object has one or more associated features. Features of an object are grouped together according to the following categories: visual, semantic and media. Multiple abstraction levels of features can be defined. For a given descriptor, the image DS also provides a link to the external descriptor extraction code and the descriptor similarity code.

Smith and Li [90] proposed a Space and Frequency View Description Scheme (SFV-DS), which provided a standard way of describing locations, sizes, regions, tilings and hierarchical decompositions of image, video and audio content in the space and the frequency domains [89]. SFV-DS offered an interface between multimedia applications and data compression and storage formats [87]. It provided a way to specify regions in space, time, frequency and resolution in terms of space and frequency views. This DS was built upon generic descriptors composed by feature vectors, and enables applications to exercise trade-off in query precision and query response time when interpreting and comparing feature vectors [91]. SFV-DS also provided a way to index views by using a Space and Frequency Graph (SFGraph). It handled details concerning access and relationship of views with different resolutions, spatial locations and sizes. It allowed a more rigorous definition of terms such as “half-resolution,” “upper-right quadrant,” or “high-pass band,” when referring to views of an image.

A graph-based description scheme was proposed by Caron et al. [12]. In this proposal, each image was represented by a graph which had a unique id. In addition, each graph was represented as a set of vertices and edges along with attributes. A vertex is denoted by a vertex ID, which is a unique integer for each vertex in the database, a vertex label and a set of attributes. An edge is labeled with an edge ID, which is a unique integer in the database, and attributes assigned to the corresponding edge. A graph was further used to illustrate such a description scheme.

Another description scheme was proposed to describe the synthetic visual content by Ostermann et al. in [64]. This synthetic visual DS consisted of several basic components: objects, animation streams and object feature structures. Each synthetic audiovisual description included synthetic audio or visual objects. Each synthetic visual description contained one or more synthetic visual objects which may have an associated animation stream. Each object had one or more associated features, and could accommodate any number of features in a modular and extensible way. Features of an object were grouped together according to the following categories: visual, semantic and media. In synthetic visual DS, each feature of an object had one or more associated descriptors. For a given descriptor, synthetic visual DS provided a link to the external descriptor extraction code and the descriptor similarity code.

5. Description Escription Definition Language

The Description Definition Language (DDL) [58] is a language that allows the creation and the representation of new description schemes and, possibly, new descriptors. It should allow the extension and modification of existing description schemes and a unique identification of all descriptors and description schemes [59]. Both primitive data types (such as text, integer, real, date, time/time index, version, etc.) and composite data types (such as histograms, graphs, RGB values, etc.) should be provided. It should also offer a mechanism to relate descriptors to data of multiple media types. DDL shall supply a rich model for links and references between one or more descriptions and the underlying data so that the relationship between the descriptions and data can be conveniently expressed. Some interesting proposals submitted to MPEG-7 are reviewed below.

Puri et al. [71] presented a MultiMedia Language (MML) which addressed the representation (for browsing, search, retrieval) and description definition problem as well as the presentation and authoring (inverse) problem. MML was primarily designed for the structured yet flexible representation of multimedia scenes. The proposed MML, derived from XML, is a superset of the Synchronized Multimedia Integration Language (SMIL) [93].

Douglass [28] proposed a DDL scheme based on formal knowledge representation languages. The object-oriented Open Knowledge Base Connectivity (OKBC) [66] model provided a language to specify meta-data schemes and descriptions required by MPEG-7. OKBC could support formal representations of content related knowledge, thus enabling emerging applications such as knowledge-based inference and

information exchange between intelligent agents. Such a formal language model captured both the semantics and the syntax of meta-data descriptions. Knowledge languages, such as OKBC, can handle the MPEG-7 application requirements for defining description schemes and descriptors related to database search and real-time ltering in a straightforward manner. Moreover, they allow the definition of MPEG-7 descriptions to be rich and rigorous enough to support intelligent multimedia applications that are emerging in parallel with the MPEG-7 standard.

ACTS-DICEMAN's DDL proposal [2] was designed as a domain-, application-, and media-independent definition language based on XML. The goal was to provide an expressive and powerful model and syntax for the expression of description schemes on audiovisual contents. It was intended to be domain and media type independent, and could be used for the description of any data (e.g. audio, video, audiovisual or text). It could also be used for the expression of descriptions. It avoided the need to define a new language for the expression of the description and made parsing easier. When used for this purpose, this DDL provided a default mechanism for linking descriptions to audiovisual material. This mechanism was applicable to still images, video and audio and potentially extensible to any multimedia application. Moreover, since DICEMAN DDL relied on XML, it was compliant with the norms used for digital texts.

Hunter's schema [37] took the optimal capabilities of the Resource Description Framework (RDF) [75], XML [30], Schema for Object-oriented XML (SOX) [83] and Document Content Description (DCD) [27], and combined and extended them, wherever needed, to satisfy all of the DDL requirements. The result was a schema based on a model consisting of classes, properties and relations between classes. It extended the model of RDF classes and properties with the addition of relations, temporal and spatial specifications and powerful data typing capabilities. These features enabled the definition of generic (temporal, spatial, conceptual) relationships and associated constraints on the attribute values of related classes. In spite of enhanced capabilities, the schema maintained its simplicity and human- and machine-readability. Since it was based on the object-oriented concept, it can easily be modified or extended.

Lennon and Wan [47] presented a Dynamic Description Framework (DDF), which provided a data model, an Application Programming Interface (API) and a serialization syntax in the description of content with respect to particular audiovisual resources. DDF attempted to incorporate benefits of declarative description of image contents with procedural methods for the creation and processing of descriptions and components of descriptions. DDF provided dynamic (procedural) creation of descriptions and components of descriptions from resources. Processing of components of descriptions and a means for linking between components of descriptions was considered in the scheme.

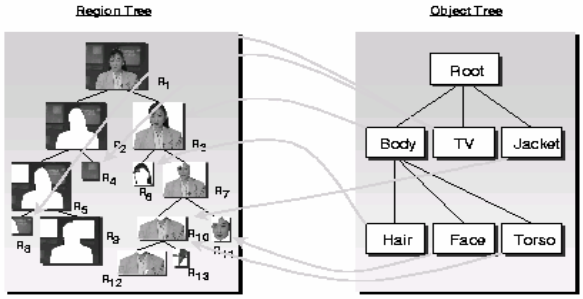

Figure 5. Anexample of reference from the Object Tree to the Region Tree.

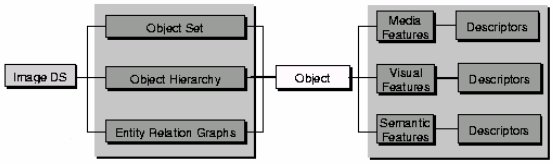

Figure 6. The structure of obect-based image DS

6. Future Trends

Based on the above discussion, it is desirable to have an intelligent database management system with three core components: multimedia data organization, information indexing and interactive query, as illustrated in Fig. 7. The lowest level of the proposed system is multimedia data organization, which analyzes and classifies collected data according to both low-level visual features and high-level semantic meanings. Image classification techniques discussed in Section 3 can be used in this stage to achieve efficient classification results. For multimedia contents belonging to different categories or classes, appropriate descriptors will be used to describe distinct characteris-tics. Based on classification and organization results, a hierarchical composite indexing structure is utilized to describe the multimedia content with multiple descriptors. The query stage should offer the user's feedback processing functions to refine the retrieval results interactively. Most current work in content-based image retrieval can be regarded as subsystems or

sub-DS in this proposed framework.

As we enter an era of information explosion, a large amount of information is created daily from places all over the world. The information may be represented in various media formats such as alphanumerical data, still pictures, graphics, 3D models, audio, speech, video, and so on. Among them, audiovisual data play a critical role since they convey the most vivid information to depict a virtual world. Future databases are expected to cover all kinds of media types to form multimedia databases. Furthermore, through the connection of the Internet, distributed multimedia databases will be linked to form a huge global database. This new infrastructure will definitely revolutionize the way how people access information [48], [46]. Emerging research and development areas include digital library, multimedia dissemination and retrieval, distributed database management, etc. Applications in the digital library and the multimedia information management are positioned to be the major driving force for the Next Generation Internet (NGI) initiative (http://www.ngi.gov). The demand to explore multimedia functionalities in Internet applications will increase rapidly in coming years [17]. Thus, the trend is to search the Web for multimedia information of interest. For example, a digital library of cultural and historic contents allows a much broader access via World Wide Web (WWW) [55]. Novel Internet technologies (such as streaming protocols and compression techniques) are also needed for efficient and cost-effective distribution of multimedia data [104]. To make all multimedia data shareable, accessible and searchable, we need a common way to describe the contents. MPEG-7 aims to create such a standard that will support these operational requirements [58], [59]. The scope of MPEG-7 is to define the standard description, but leave open specific ways of feature extraction and multimedia search. The structure of a typical multimedia database accessible via the Internet is shown in Fig. 8, where the description of all data should be consistent to the MPEG-7 standard so that data access are interoperable. Thus, future development of multimedia information management will be strongly influenced by the MPEG-7 standard and tied with Internet applications.

In the following, we identify four research areas worth thorough study in the near future.

(1). Multimedia Database Design:

A good database abstraction model should reect real world problems with a simple structure and an easy implementation. The relational database model has been proved to be successful in handling traditional alphanumeric data. However, it is difficult to be extended to the context of multimedia data. It is not clear what serves as a good multimedia database model even though there has been effort along this direction.

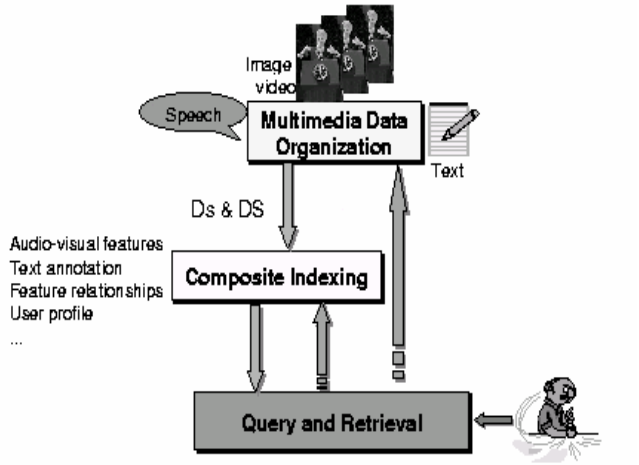

Figure 7. The architecture of a three-stage multimedia information retrieval system

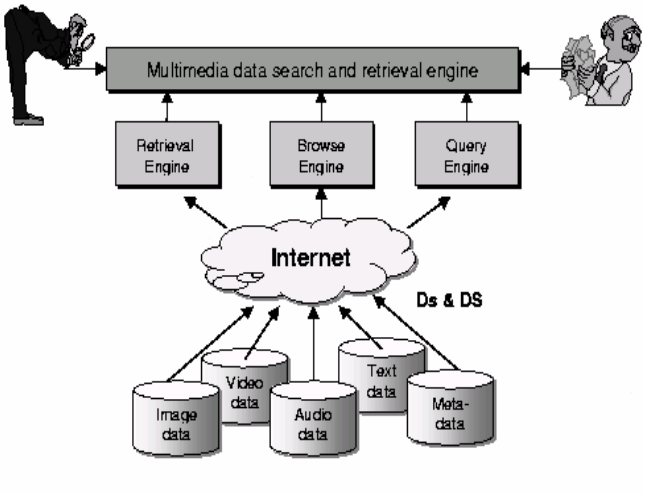

Figure 8. Multimedia database and retrieval over the Internet

(2). Intelligent Information Access:

Current techniques for information access on the Web are far from maturity and have much room for major improvements. Search engines neither offer comprehensive indices of the Web nor provide multimedia data query and retrieval. There exists a great difficulty in accurately ranking relevance results and rapidly processing human's feedback. It is difficult to predict what the next generation Internet multimedia search engine look like at this point. Without a high performance search engine, large information collections

available over the World Wide Web are actually of little value to most users.

(3). Multimodal Human-Computer Interface:

The human-machine interface is another important issue in the next generation Web-based multimedia database applications. Recent advances in signal processing technologies, coupled with an explosion of information [84], has given rise to a number of novel human-computer interaction modalities such as speech, vision-based gesture recognition, image understanding, eye tracking, etc. Human interactivity and perception is heavily involved in the multimedia information access process. How to utilize user's feedback and involvement to achieve fast information access and accurate retrieval is very crucial to multimedia information management. Since real world databases contain a hybrid of various media types including image, audio, video, and texts. Multi-modal human-computer interface design will be a promising direction in the near future.

(4). Data Mining:

Data mining, also known as knowledge discovery in the database discipline, gives organizations tools to sift through these vast data stores to find trends, patterns, and correlations that can guide strategic decision making [33]. The sharing of a huge multimedia database through the Internet will make more information available. Sifting through the growing mountain of Web data demands an increasingly discerning search engine, one that can reliably assess the quality of sites, not just their relevance [16]. Also, user's involvement and frequent transaction in Internet databases provide valuable patterns and trends. In the future, data mining will deal with not only traditional alphanumeric data but also image, video, speech and audio data. The corresponding multimedia data mining tools should help to discover the hidden knowledge inside Web data and provide intelligent decision making.

7. Conclusion

In this work, state-of-the-art technologies of image content analysis, indexing and retrieval for image database management were reviewed. By following the MPEG-7 standard activities, we examined important technologies of feature extraction, visual feature descriptors and description schemes. Future trends were predicted while promising research and development directions were suggested.

Acknowledgement

This research has been funded in part by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center with additional support from the Annenberg Center for Communication at the University of Southern California and the California Trade and Commerce Agency.

Reference

[1] Abdel-Mottaleb M., Krishnamachari, S., “Color representation by multiple local histogram,” *ISO/IEC/JTC1/ SC29/WG11*, pp. 648, Lancaster, UK, Feb. (1999).

[2] *ACTS-DICEMAN*, *ISO/IEC/JTC1/SC29/ WG11*, pp. 184, Lancaster, UK, Feb. (1999).

[3] *ACTS-DICEMAN*, *ISO/IEC/JTC1/SC29/ WG11*, pp. 186, Lancaster, UK, Feb. (1999).

[4] Aksoy, S. and Haralick, R. M., “Textural features for image database retrieval,” in IEEE CVPR'98 Workshop on Content-Based Access of Image and Video Libraries, (1998).

[5] Alata, O., Cariou, C., Ramannanjarasoa, C. and Najim, M., “Classification of rotated and scaled textures using HMHV spectrum estimation and the Fourier-Mellin Transform,” in IEEE International Conference on Image Processing, (1998).

[6] Bach, J. R., Fuller, C., Gupta, A., Hampapur, A., Horowitz, B., Humphrey, R., Jain, R., and Shu, C. F., “The Virage image search engine: an open framework for image management,” in SPIE Storage and Retrieval for Image and Video Databases V, (1996).

[7] Beek, P. V., Qian, R. and Sezan, I., “Scalable blob histogram descriptor,” *ISO/IEC/JTC1/SC29/WG11*, pp. 430, Lancaster, UK, Feb. (1999).

[8] Bhanu, B., Qing, S. and Peng, J., “Learning integrated online indexing for image databases,” in IEEE International Conference on Image Processing, (1998).

[9] Bober, M., ISO/IEC/JTC1/SC29/WG11 ,pp. 320, Lancaster, UK, Feb. (1999).

[10] Bouet, M. and Djeraba, C., “Powerful image organization in visual retrieval systems,” in *ACM Multimedia'98*, Bristol, UK, (1998).

[11] Bowman, C., Danzig, P., Hardy, D., Maber, U., and Schwartz, M., “The Harvest

information discovery and access system," in *the 2nd International World Wide Web Conference*, (1994).

[12] Caron, C., Challapali, K. and Yan, Y., ISO/IEC/JTC1/SC29/WG11, pp. 646, Lancaster, UK, Feb. (1999).

[13] Carson, C., Belongie, S., Greenspan, H. and Malik, J., "Region-based image querying," in *IEEE CVPR'97 Workshop on Content-Based Access of Image and Video Libraries*, (1997).

[14] Carson, C., Belongie, S., Greenspan, H., and Malik, J., "Blobworld: image segmentation using expectation- maximization and its application to image querying," in submitted to *IEEE Trans. on Pattern Analysis and Machine Intelligence*, (1998).

[15] Cha, K. H., Lee, H. A., Park, J. D., Ryu P.-M., Chae Y.-S., and Park S.-Y., "Representation of the situational meaning of an image based on domain ontology," ISO/IEC/JTC1/SC29/WG11 pp. 331, Lancaster, UK, Feb. (1999).

[16] Chakrabarti, S., Dom, B., Kumar, S., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D. and Kleinberg, J., "Mining the Web's link structure," in *IEEE Computer Magazine*, Aug. (1999).

[17] Chang, S. F., Smith, J. R. and Meng, H. J., "Exploring image functionalities in WWW applications - development of image/video search and editing engines", in *IEEE International Conference on Image Processing*, (1997).

[18] Chang, T. and Kuok, C. C. J., "Texture analysis and classification with tree-structured wavelet transform," in *IEEE Trans. on Image Processing*, (1993).

[19] Cieplinski, L., *ISO/IEC/JTC1/SC29/ WG11/*, Lancaster, UK, Feb. pp. 319, (1999).

[20] "Core experiments on MPEG-7 color and texture descriptors," *ISO/IEC/JTC1/ SC29/WG11* Seoul, Korea, Mar. pp. 2691, (1999).

[21] Cutting, D., Karger, D., Pedersen, J., and Tukey, J., "Scatter/gather: A cluster-based approach to browsing large document collections," in A*CM SIGIR'92*, (1992).

[22] Deng, Y., Kenney, C., Moore, M. S. and Manjunath, B. S., "Peer group flltering and perceptual color quantization", in *IEEE International Symposium on Circuits and Systems*, (1999).

[23] Deng, Y., Manjunath, B. S., Shin, H. and Choi, Y., "A color descriptor for MPEG-7: variable-bin color histogram," *ISO/IEC JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 76 (1999).

[24] "Description of core experiments for MPEG-7 motion/shape", *ISO/IEC JTC1/SC29/WG11N2690*, Seoul, Korea, Mar. (1999).

[25] Dimitrova, N., Agnihotri, L., Martino, J. and Elenbaas, H., "Super-histogram for video classification and program," *ISO/IEC/JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 641 (1999).

[26] Dimitrova, N., McGee, T., Elenbaas, H. and Martino, J., "Video content management in consumer devices," in *IEEE Trans. on Knowledge and Data Engineering*, Nov. (1998).

[27] "Document content description (DCD) for XML", Submission to *W3C*, Jul. (1998).

[28] Douglass, R. J., "Description definition language (DDL), knowledge representation language for MPEG-7 DDL," *ISO/IEC JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 124 (1999).

[29] Dowe, J., "Content-based retrieval in multimedia imaging," in *SPIE Storage and Retrieval for Image and Video Databases II*, (1993).

[30] "Extensible markup language (XML) 1.0," *REC-xml-(1998)0210, W3C Recommendation*, Feb. (1998).

[31] Ferman, A. M. and Tekalp, A. M., "Efficient filtering and clustering methods for temporal video segmentation and visual summarization," in *Journal of Visual Communication and Image Representation*, Vol. 9,No. 4, pp. 336-351, (1998).

[32] Ferman, A. M., Tekalp, A. M., Mehrotra, R., "Histogram-based color descriptors for multiple frame color characterization," *ISO/IEC/JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 529 (1999).

[33] Ganti, V., Gebrke, J. and Ramakrishnan, R., "Mining very large database," in *IEEE Computer Magazine*, Aug. (1999).

[34] Gonzalez, R. C., Woods, R. E., *Digital Image Processing*, Addison Wesley Publishing Company, (1993).

[35] Huang, J., Kumar, S. R. and Mitra, M., "Combining supervised learning with color correlograms for content-based image retrieval," in *ACM Multimedia'97*, Seattle, WA, (1997).

[36] Huang, J., Kumar, S. R. and Zabih, R., "An automatic hierarchical image classification scheme," in *ACM Multimedia'98*, Bristol, UK,

(1998).
[37] Hunter, J., "A proposal for an MPEG-7 description definition language (DDL) ," *ISO/IEC/JTC/SC29/WG11*, Lancaster, UK, Feb. pp. 547 (1999).
[38] "IBM Almaden Research Center, Technical summary of color descriptors for MPEG-7," *ISO/IEC JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 165 (1999).
[39] Iwayama, M. and Tokunaga, T., "Cluster-based text categorization: A compression of category search strategies," in *ACM SIGIR'95*, (1995).
[40] Jung, S., Kim, K., Chun, B. T., Lee, J. Y. and Bae, Y., "Color descriptor by using picture information measure of subregions in video sequence," *ISO/IEC JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 549 (1999).
[41] Khotanzad, A. and Hong, Y. H., "Invariant image recognition by Zernike moments," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, May (1990).
[42] Kim, H. J., Lee, J. S., Jun, S. B., Song, J. M., Lee, H. Y., "Descriptor for quantized color using HMMD color model," *ISO/IEC/JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 669 (1999).
[43] Kim, J. D. and Kim, H. K., "Shape descriptor based on multi-layer eigen vector," *ISO/IEC/JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 517 (1999).
[44] Kim, W. Y. and Kim, Y. S., "A rotation invariant geometric shape descriptor using Zernike moment," *ISO/IEC JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 687 (1999).
[45] Kim, Y.-S. and Kim, W.-Y., "Content-based trademark retrieval system using visually salient feature," in *Journal of Image and Vision Computing*, Aug. (1998).
[46] Lawrance, S. and Giles, C. L., "Searching the Web: general and scientific information access," in *IEEE Communication Magazine*, Jan. (1999).
[47] Lennon, A. and Wan, E., "Dynamic description framework," *ISO/IEC /JTC/SC29/WG11*, Lancaster, UK, Feb. pp. 487 (1999).
[48] Li, C. S. and Stone, H. S., "Digital library using next generation internet," in *IEEE Commu-nication Magazine*, Jan. (1999).
[49] Lipson, P., Grimson, W. E. L. and Sinha, P., "Configuration based scene classification and image indexing," in *IEEE Conference on Computer Vision and Pattern Recognition*, (1997).
[50] Liu, F. and Picard, R., "Periodicity, directionality and randomness: Wold features for image modeling and retrieval," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 7, Jul. (1996).
[51] Mahmood, T. S., "Location hashing: an efficient indexing method for object queries in image databases," *ISO/IEC JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 144 (1999).
[52] Manjunath, B. S. and Ma, W., "Texture features for browsing and retrieval of image data," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Aug. (1996).
[53] Medasani, S. and Krishnapuram, R., "Categorization of image databases for efficient retrieval using robust mixture decomposition," in *IEEE CVPR'98 Workshop on Content-Based Access of Images and Video Libraries*, (1998).
[54] Minka, T. P. and Picard, R. W., "Interactive learning with a `society of models", in *Pattern Recognition, 30(4)*, pp. 565-581, Apr. (1997).
[55] Mintzer, F., "Developing digital library of cultural content for Internet access," in *IEEE Communication Magazine*, Jan. (1999).
[56] Mottaleb, M. A., "A descriptor for the edges in still images," *ISO/IEC /JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 649 (1999).
[57] "MPEG-7 application document V.8," *ISO/IEC JTC1/SC29/WG11 N2728*, Seoul, Korea, Mar. (1999).
[58] "MPEG-7 Context, objectives and technical roadmap V.11," *ISO/IEC JTC1/SC29/WG11 N2729*, Seoul, Korea, Mar. (1999).
[59] "MPEG-7 requirements document V.8," *ISO/IEC JTC1/SC29/WG11 N2727*, Seoul, Korea, Mar. (1999).
[60] Muller, K. and Ohm, J. R., "Wavelet-based contour descriptor," *ISO/IEC/JTC1/ SC29/WG11*, Lancaster, UK, Feb. pp. 567 (1999).
[61] Nevel, A. V., "Texture synthesis via matching first and second order statistics of a wavelet frame decomposition," in *IEEE International Conference on Image Processing*, (1998).
[62] Niblack, W., Berber, R., Equitz, W., "Flickner M., Glasman E., Petkovic D., and Yanker P., The QBIC project: querying images by content using color, texture and shape", in *SPIE Storage and Retrieval for Image and Video Database II*, Feb. (1993).
[63] Niblack, W., Zhu, X., Hafner, J., Breuel, T., Pondeleon, D., Petkovic, D., Flickner, M., Upfae, L., Nin, S., Sull, S., Dom, B., Yeo, B. L., "Srinivasan S., Zivkovic D., and Penner M., Updates to the QBIC system," in *SPIE Storage and Retrieval for Image and Video Database VI*, Jan. (1998).

[64] Ohm, J. R., Makai, B., ISO/IEC /JTC1/ SC29/WG11 Lancaster, UK, Feb. pp. 563-564 (1999).
[65] Ohm, J. R. and Bunjamin, F., *ISO/IEC JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 566 (1999).
[66] "Open knowledge base connectivity home page,"
[67] Ostermann, J., Rajendran, R. K., Puri, A., Huang, Q., *ISO/IEC/JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 472 (1999).
[68] Paek, S., Benitez, A., Chang, S. F., Li, C. S., Smith, J. R., Bergman, L. D., Puri, A., Swain, C. and Ostermann, J., *ISO/IEC/JTC1/ SC29/ WG11*, Lancaster, UK, Feb. pp. 480 (1999).
[69] Paul, E., Vet, V. and Mars, N. J., "Bottom-up construction of ontologies," in *IEEE Trans. on Knowledge and Data Engineering*, (1998).
[70] Pratt, W. K., *Digital Image Processing*, 2nd edition, a Wiley-Interscience Publication, (1991).
[71] Puri, A., Huang, Q., Smith, J. R., Kim, M. C., Mohan, R., Li, C. S., Bergman, L. D., Eleftheriadis, A., Benetiz, A. B., Fang, Y., Rajendran, R. K. and Chang, S. F., "MPEG multimedia language (MML): a proposal for MPEG-7 DDL," *ISO/IEC JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 484 (1999).
[72] Rao, A. R. and Lohse, G. L., "Towards a texture naming system identifying relevant dimensions of texture," in *IEEE Conference on Visualization*, Oct. (1993).
[73] Ratan, A. L. and Grimson, W. E. L., "Training templates for scene classification using a few examples," in *IEEE CVPR'97 Workshop on Content-Based Access of Image and Video Libraries*, (1997).
[74] "Report of the ad-hoc group on MPEG-7 evaluation logistics," ISO/IEC/JTC1/SC29/WG11/ MPEG99/ Wxxxx, Lancaster, UK, Feb. (1999).
[75] "Resource description framework (RDF) schema specification," *WD-rdf-schema-(1998)1030*, W3C Working Draft, Oct. (1998).
[76] Ricoh Company Ltd., "MINDS's descriptors for still images - spatial edge distribution descriptor," *ISO/IEC/JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 102 (1999).
[77] Ricoh Company Ltd., "MINDS's descriptors for still images - spatial texture distribution descriptor," *ISO/IEC JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 104 (1999).
[78] Ro, Y. M., Kim, S. Y., You, K. W., Kim, M. and Kim, J., "Texture description using atoms of matching pursuits," *ISO/IEC JTC1/SC29/ WG11*, Lancaster, UK, Feb. pp. 612 (1999).
[79] Ro, Y. M., "Matching pursuit: contents featuring and image indexing," in *SPIE Multimedia storage and archiving system III*, (1998).
[80] Rui, Y., Huang, T. and Mehrotra, S., "Content-based image retrieval with relevance feedback in MARS," in *IEEE International Conference on Image Processing*, pp. 815-818, Oct. (1997).
[81] Rui, Y., Huang, T. S. and Chang, S. F., "Image retrieval: current techniques, promising directions, and open issues," in *Journal of Visual Communication and Image Representation*, Vol. 10, No. 1, Mar. (1999).
[82] Salton, G. and Araya, J., "On the use of clustered file organization in information search and retrieval," in *Tech. Rep. TR89-989*, Department of Computer Science, Cornell University, (1989).
[83] "Schema for object-oriented XML (SOX) ," *NOTE-SOX-(1998)0930*, Submission to W3C, Sep. (1998).
[84] Sharma, R., Pavlovic, V. I. and Huang, T. S., "Toward multimodal human-computer interface," in *Proceedings of the IEEE*, May (1998).
[85] Silberschatz, A., Korth, H. F. and Sudarshan, S., *Database System Concepts*, 3rd Edition, McGraw-Hill Publisher, (1996).
[86] Simoncelli, E. P. and Portilla, J., "Texture characterization via joint statistics of wavelet coefficient magnitudes," in *IEEE International Conference on Image Processing*, (1998).
[87] Smith, J. R. and Castelli, V. and Li, C. S., "Adaptive storage and retrieval of large compressed images," in *SPIE Storage and Retrieval for Image and Video Databases VII*, Jan. (1999).
[88] Smith, J. R. and Chang, S. F., "An image and video search engine for the World-Wide Web," in *ACM Multimedia'96*, (1996).
[89] Smith, J. R. and Chang, S.F., "Joint adaptive space and frequency basis selection," in *IEEE International Conference on Image Processing*, Oct. (1997).
[90] Smith, J. R. and Li, C. S., *ISO/IEC JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 483 (1999).
[91] Smith, J. R., "Query vector projection access method," in *SPIE Storage and Retrieval for Image and Video Databases VII*, Jan. (1999).
[92] "Squire D. M., Learning a similarity-based distance measure for image database organization from human partitionings of an image set," in *SPIE Multimedia Storage and*

Archiving System III, (1998).

[93] "Synchronized multimedia integration language (SMIL) ," WD-smil-0202, W3C Working Draft, Feb. (1998).

[94] Szummer, M. and Picard, R. W., "Indoor-outdoor image classification," in *IEEE CVPR'98 Workshop on Content-Based Access of Image and Video Libraries*, (1998).

[95] Tabatabai, A., "Color representation for visual objects," *ISO/IEC JTC1/SC29/ WG11*, Lancaster, UK, Feb. pp. 576 (1999).

[96] Tamura, H., Mori, S. and Yamawaki, T., "Textural features corresponding to visual perception," in *IEEE Trans. on Sys. Man., and Cyber.*, Jun. 1978.

[97] Tao, B. and Dickinson, B. W., "Recognition and retrieval of textured images using gradient indexing," in *IEEE International Conference on Image Processing*, (1998).

[98] Tektronix Inc., "Normalized contour as a shape descriptor for visual objects," *ISO/IEC/JTC1/SC29/WG11*, Lancaster, UK, Feb. pp. 579 (1999).

[99] Vellaikal, A. and Kuo, C. C. J., "Hierarchical clustering techniques for image database organization and summarization," in *SPIE Multimedia Storage and Archiving System III*, Nov. (1998).

[100] Wan, X. and Kuo, C. C. J., "Image retrieval based on JPEG compressed data," in *SPIE Multimedia Storage and Archiving Systems*, Nov. (1996).

[101] Weiss, R., Velez, B., Sheldon, M., Namprempre, C., Szilagyi, P., Duda, A., and Gifford, D., "Hy-Pursuit: A hierarchical network search engine that exploits content-link hypertext clustering," in *ACM Hypertext'96*, (1996).

[102] Won, C. S., et al, "Efficient color feature extraction in compressed video," in *SPIE Storage and Retrieval for Image and Video Databases VIII*, (1999).

[103] Won, C. S., Park, D. K., Yoo, S. J., Park, S. J., "Generalized image histogram," ISO/ IEC/JTC1/SC29/WG11, Lancaster, UK, Feb. pp. 324 (1999).

[104] Wong, S. and Tjandra, D., "A digital library for biomedical imaging on the Internet," in *IEEE Communication Magazine*, Jan. (1999).

[105] Wood, M., Campbell, N. and Thomas, B., "Iterative refinement by relevance feedback in content-based digital image retrieval," in *ACM Multimedia'98*, Bristol, UK, (1998).

[106] Wu, P., Ma, W.Y., Manjunath, B. S., Shin, H. and Choi, Y., "Texture descriptor," ISO/IEC/JTC1/SC29/WG11, Lancaster, UK, Feb. pp. 77 (1999).

[107] Lamdan, Y. and Wolfson, H. J., "Geometric hashing: a general and efficient model-based recognition scheme," in *IEEE Conference on Computer Vision and Pattern Recognition*, (1988).

[108] Yang, Z. and Kuo, C. C. J., "A semantic classification and composite indexing approach to robust image retrieval," in *IEEE International Conference on Image Processing*, Oct. (1999).

[109] Yang, Z. and Kuo, C. C. J., "Content- based image retrieval via adaptive multi-feature templates," in *SPIE Multimedia Storage and Archiving System IV*, Sep. (1999).

[110] Yang, Z. and Kuo, C. C. J., "Intelligent image database indexing and query system," in *SPIE Application of Digital Image Processing XXII*, Jul. (1999).

[111] Yang, Z., Wan, X., and Kuo, C. C. J., "Interactive image retrieval: concept, procedure and tools," in *IEEE 32nd Asilomar Conference*, Montery, CA, Nov. (1998).

[112] Yu, H. H. and Wolf, W., "Scenic classification methods for image and video databases," in *SPIE Digital Image Storage and Archiving Systems*, (1995).

Accepted: Sept. 5, 1999