IRM PRESS
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.irm-press.com

ITB13850

This chapter appears in the book, ***Semantic-Based Visual Information Retrieval***
by **Y.J. Zhang** © 2007, Idea Group Inc.

Chapter I

Toward High-Level Visual Information Retrieval

Yu-Jin Zhang, Tsinghua University, Beijing, China

Abstract

Content-based visual information retrieval (CBVIR) as a new generation (with new concepts, techniques, mechanisms, etc.) of visual information retrieval has attracted many interests from the database community. The research starts by using a low-level feature from more than a dozen years ago. The current focus has shifted to capture high-level semantics of visual information. This chapter will convey the research from the feature level to the semantic level by treating the problem of semantic gap under the general framework of CBVIR. This high-level research is the so-called semantic-based visual information retrieval (SBVIR). This chapter first shows some statistics about the research publications on semantic-based retrieval in recent years; it then presents some existing approaches based on multi-level image retrieval and multi-level video retrieval. It also gives an overview of several current centers of attention by summarizing certain results on subjects such as image and video

annotation, human-computer interaction, models and tools for semantic retrieval, and miscellaneous techniques in application. Before finishing, some future research directions, such as domain knowledge and learning, relevance feedback and association feedback, as well as research at even a high level such as cognitive level, are pointed out.

Introduction

It is said that "a picture is worth a thousand words." Human beings obtain the majority of information from the real world by visual sense. This could include all entities that can be visualized, such as image and video (a chain/sequence of images) in a narrow sense, as well as animation, charts, drawings, graphs, multi-dimensional signals, text (in fact, many documents are used in image form, as indicated by Doermann, 1998), and so forth in a more general sense.

With the fast technique progress of computer science, electronics, medium capturing, and so forth, and the rapidly rising use of the Internet and the growing capability of data storage, the quantity of visual information expands dramatically and results in many huge visual information databases. In addition, many data are created and collected by amateurs, which is quite different than by professional people (Luo, Boutell, & Brown, 2006). In addition, visual media become a widespread information format in the World Wide Web (WWW) in which data are dispersed in various locations. All these make the search of required visual information more complex and time-consuming (Zhang, 2006). Along with the quickly increasing demands to create and store visual information comes the need for a richer set of search facilities. Providing tools for effective access, retrieval, and management of huge visual information data, especially images and videos, has attracted significant research efforts. Several generations of techniques and systems have been developed.

Traditionally, textual features such as captions, file names, and especially keywords have been used in searching required visual information. However, the use of keywords in the search is not only cumbersome but also inadequate to represent the riche content of visual information. Images are snapshots of the real world. Due to the complexity of scene content, there are many images for which no words can exactly express their implications. Image is beyond words, so it has to be seen and must be searched as image by content (i.e., object, purpose, scene, style, subject, etc.).

Content-based visual information retrieval has attracted many interests, from image engineering, computer vision, and database community. A large number of researches, especially on feature-based techniques, have been developed and have achieved plentiful and substantial results (Bimbo, 1999; Rui, Huang, & Chang, 1999; Smeulders et al., 2000; Zhang, 2003). However, in light of the complexity of the real world, low-level perceptive cues/indexes are not enough to provide suitable interpretation. To probe further, some higher-level researches and techniques for content understanding are mandatory. Among three broad categories of high-level techniques—synthetic, semantic, and semiotic—the semantic approach is quite natural from the understanding point of view. Nevertheless, from feature to semantic, there is a semantic gap. Solving this problem has been a focal point in

content-based visual information retrieval. This chapter will summarize some recent research results and promote several new directions in higher-level researches.

Background

In the following, visual information will refer mainly to image and video, although other media also will be covered with some specific considerations.

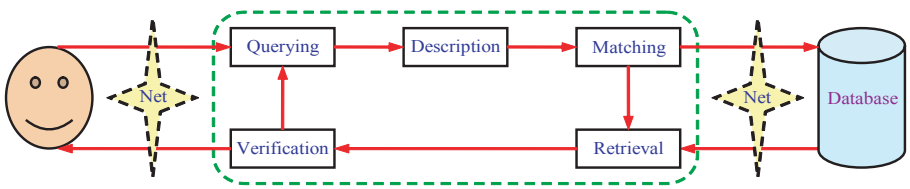
Content-Based Visual Information Retrieval

More than 10 years ago, the term *content-based image retrieval* made its first appearance (Kato, 1992). Content-based image retrieval (CBIR) could be described as a process framework for efficiently retrieving images from a collection by similarity. The retrieval relies on extracting the appropriate characteristic quantities describing the desired contents of images. Shortly thereafter, content-based video retrieval (CBVR) also made its appearance in treating video in similar means as CBIR-treating images. Content-based visual information retrieval (CBVIR) soon combines CBIR and CBVR together. Also, the MPEG-7 standard was initiated with the intention to allow for efficient searching, indexing, filtering, and accessing of multimedia (especially image and video) contents (Zhang, 2003).

A general scheme for content-based visual information retrieval is shown in Figure 1. The system for retrieval is located between the information user and the information database. It consists of five modules: Querying, for supporting the user to make an inquiry; Description, for capturing the essential content/meaning of inquest and transferring it to internal representation; Matching, for searching required information in a database; Retrieval, for extracting required information from a database; and Verification, for ensuring/confirming the retrieval results.

Research on CBVIR today is a lively discipline that is expanding in breadth. Numerous papers have appeared in the literature (see the next section), and several monographs have been published, such as Bimbo (1999) and Zhang (2003). As happens during the maturation process of many disciplines, after early successes in a few applications, research then concentrates on deeper problems, challenging the hard problems at the crossroads of the

Figure 1. General scheme for content-based visual information retrieval



discipline from which it was born: image processing, image analysis, image understanding, databases, and information retrieval.

Feature-Based Visual Information Retrieval

Early CBVIR approaches often rely on the low-level visual features of image and video, such as color, texture, shape, spatial relation, and motion. These features are used widely in the description module of Figure 1 for representing and describing the contents. It is generally assumed that the perceptual similarity indicates the content similarity. Such techniques are called feature-based techniques in visual information retrieval.

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. Texture refers to the visual patterns that have properties of homogeneity that do not result from the presence of only a single color or intensity. It is an innate property of virtually all surfaces, including clouds, trees, bricks, hair, and fabric. The shape feature is at some higher level than that of color and texture, as the focus is now on interesting objects. Some required properties for shape features are translation, rotation, and scaling invariants. Once interesting objects are determined, their spatial relationship will provide more information about the significance of the whole scene, such as the structural arrangement of object surfaces and the association with the surrounding environment.

The aforementioned features often are combined in order to provide more complete coverage of various aspects of image and video. For example, the global color feature is simple to calculate and can provide reasonable discriminating power, but using the color layout feature (both color feature and spatial relation feature) is a better solution to reduce false positives when treating large databases. A number of feature-based techniques and examples (using single features and/or composite features) can be found in Zhang (2003).

There are several early surveys on this subject. In Rui et al. (1999), more than 100 papers covering three fundamental bases of CBIR—image feature representation and extraction, multidimensional indexing, and system design—are reviewed. In Smeulders et al. (2000), a review of 200 references in content-based image retrieval published in the last century has been made. It starts with discussing the working conditions of content-based retrieval: patterns of use, types of pictures, the role of semantics, and the sensory gap. The discussion on feature-based retrieval is divided into two parts: (1) the features could be extracted from the pixel-level, such as color, texture, and local geometry; and (2) the features need to be extracted from the pixel-group level (more related to objects), such as accumulative and global features (of pixels), salient points, object and shape features, signs, and structural combinations.

From Features to Semantics: Semantic Gap

Although many efforts have been put on CBVIR, many techniques have been proposed, and many prototype systems have been developed, the problems in retrieving images according to image content are far from being solved. One problem of the feature-based technique

is that there is a considerable difference between users' interest in reality and the image contents described by using only the previously mentioned low-level perceptive features (Zhang, Gao, & Luo, 2004), although all current techniques assume certain mutual information between the similarity measure and the semantics of images and videos. In other words, there is a large gap between content description based on low-level features and that of human beings' understanding. As a result, these feature-based approaches often lead to unsatisfying querying results in many cases. In Smeulders et al. (2000), some discussions on the semantic gap also are presented.

One solution to fill the semantic gap is to make the retrieval system work with low-level features while the user puts in high-level knowledge so as to map low-level visual features to high-level semantics (Zhou & Huang, 2002). Two typical early methods are to optimize a query request by using relevance feedback and a semantic visual template (Chang, 1998) and to interpret progressively the content of images by using interactive interface (Castelli, Bergman, Kontoyiannis, et al., 1998).

Nowadays, the mainstream of the research converges to retrieval based on semantic meaning, which tries to extract the cognitive concept of a human by combining the low-level features in some way. However, semantic meaning extraction based on feature vectors is difficult, because feature vectors indeed cannot capture the perception of human beings. For example, when looking at a colorful image, an ordinary user hardly can figure out the color histogram from that image but rather is concerned about what particular color is contained.

The methods toward semantic image retrieval have been categorized roughly into the following classes (Cheng, Chen, Meng, Sundaram, & Zhong, 2005): (1) automatic scene classification in whole images by statistically based techniques; (2) methods for learning and propagating labels assigned by human users; (3) automatic object classification using knowledge-based or statistical techniques; (4) retrieval methods with relevance feedback during a retrieval session. An indexing algorithm may be a combination of two or more of the aforementioned classes.

Main Thrust

First, some statistics about research publication are provided and analyzed. Then, demonstrative approaches for multi-level image and video retrievals are presented. Finally, an overview of recent research works is made by a brief survey of representative papers.

Statistics about Research Publications

The study for CBVIR has been conducted for more than 10 years. As a new branch of CBVIR, SBVIR recently has attracted many research efforts. To get a rough idea about the scale and progress of research on (general) image retrieval and semantic-based image retrieval for the past years, several searches in EI Compendex database (http://www.ei.org) for papers published from 1995 through 2004 have been made. In Table 1, the results of two searches

Table 1. List of records found in the title field of EI Compendex

Searching Terms	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	Total
(1) Image Retrieval	70	89	81	135	157	166	212	237	260	388	1795
(2) Semantic Image Retrieval	0	1	1	2	4	5	5	9	12	20	59
Ratio of (2) over (1)	0	1.12	1.23	1.48	2.55	3.01	2.36	3.80	4.62	5.15	3.29

in the title field of EI Compendex for the numbers of published papers (records) are listed; one term used is *image retrieval* and other is *semantic image retrieval*. The papers found out by the second term should be a subset of the papers found out by the first term.

It is seen from Table 1 that both numbers (of papers) are increasing in that period, and the number of published papers with the term *semantic image retrieval* is just a small set of papers with the term *image retrieval*.

Other searches take the same terms as used for Table 1, but are performed in the field of title/abstract/subject. The results are shown in Table 2.

The numbers of records in Table 2 for both terms are augmented during that period in comparison with that of Table 1. This is expected, as the fields under search now also include abstract field and subject field in addition to only title field.

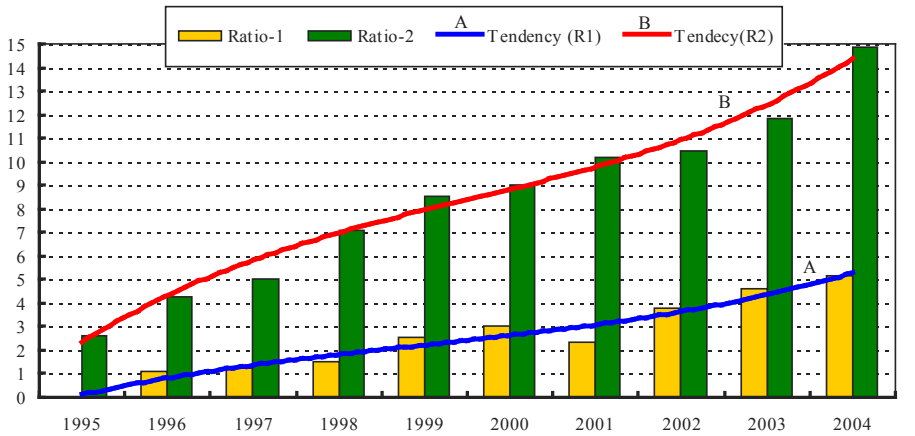
Comparing the ratios of SIR over IR for 10 years in two tables, these ratios in Table 2 are much higher than those ratios in Table 1. This difference indicates that the research for semantic retrieval is still in an early stage (many papers have not put the word *semantic* in the title of papers), but this concept starts to get numerous considerations or attract much attention (*semantic* appeared already in abstract and/or subject parts of these papers).

In order to have a closer comparison, these ratios in Table 1 and Table 2 are plotted together in Figure 2, in which light bars represent ratios from Table 1, and dark bars represent ratios

Table 2. List of records found in the subject/title/abstract field of EI Compendex

Searching Terms	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	Total
(1) Image retrieval	423	581	540	649	729	889	1122	1182	1312	2258	9685
(2) Semantic image retrieval	11	25	27	46	62	80	114	124	155	335	979
Ratio of (2) over (1)	2.60	4.30	5.00	7.09	8.50	9.00	10.16	10.49	11.81	14.84	10.11

Figure 2. Comparison of two groups of ratios



from Table 2. In addition, the tendencies of ratio developments are approximated by third-order polynomial. It is clear that many papers have the semantic concept in mind, although they do not always use the word *semantic* in the title. It is also clear that both ratios have the tendency to increase, with the second ratio going up even faster.

Multi-Level Image Retrieval

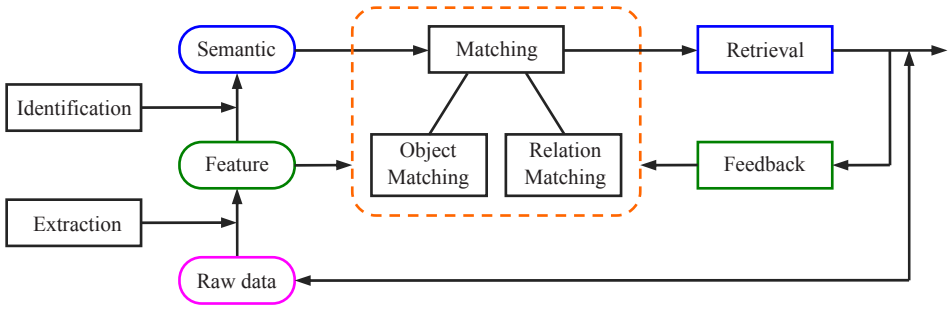
Content-based image retrieval requires capturing the content of images, which, in turn, is a challenging task. To improve the performance of image retrieval, there is a strong trend to analyze images in a hierarchical way so as to represent and describe image contents progressively toward the semantic level.

Multi-Level Representation

Many approaches have been proposed to represent and describe the content of images in a higher level than the feature level, which should be more corresponding to human beings' mechanisms for understanding and which also reflects the fuzzy characteristics of image contents. Several representation schemes have been proposed to represent the contents of images in different levels (Amir & Lindenbaum, 1998), such as the three-level content representation, including feature-level content, object-level content, and scene-level content (Hong, Wu, & Singh, 1999); and the five-level representation, including region level, perceptual region level, object part level, object level, and scene level (Jaimes & Chang, 1999).

On the other side, people distinguish three layers of abstraction when talking about image database: (1) raw data layer, (2) feature layer, and (3) semantic layer. The raw data are

Figure 3. A general paradigm for object-based image retrieval



original images in the form of pixel matrix. The feature layer shows some significant characteristics of the pixel patterns of the image. The semantic layer describes the meaning of identified object in images. Note that the semantic level also should describe the meaning of an image as a whole. Such a meaning could be obtained by the analysis of objects and the understanding of images.

According to the previous discussions, a multi-layer approach should be used in which the object characterization plays an important role. The basic idea behind such an approach is that images in each object class have similar semantic meanings and visual perceptions (Dai & Zhang, 2005). A general paradigm is shown in Figure 3 (Zhang, 2005a). Two important tasks are:

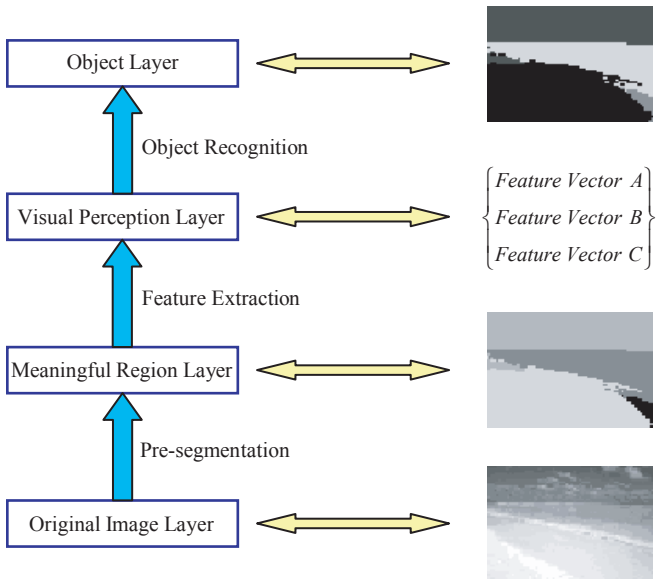
1. **Extracting meaningful regions:** In order to be able to base image retrieval on objects, the interesting regions related to objects should be extracted first. This process relates the raw data layer to the feature layer.
2. **Identification of interesting objects:** Based on the extracted regions, (perceptual) features should be taken out, and those required objects could be identified. This corresponds to the step from the feature layer to the object layer.

Multi-Level Description

A typical realization of the aforementioned general paradigm consists of four layers: original image layer, meaningful region layer, visual perception layer, and object layer, as depicted in Figure 4 (Gao, Zhang, & Merzlyakov, 2000). One note here is that instead of deriving accurate measures of object properties for further identification or classification, the primary concerns in CBIR are to separate required regions and to obtain more information related to the semantic of these regions (Zhang, 2005b).

In Figure 4, the left part shows the multi-level model, while the right part gives some presentation examples. The description for a higher layer could be generated on the basis of the description obtained from the adjacent lower layer by using the following various techniques: presegmentation from the original image layer to the meaningful region layer (Luo et al.,

Figure 4. Multi-level image description



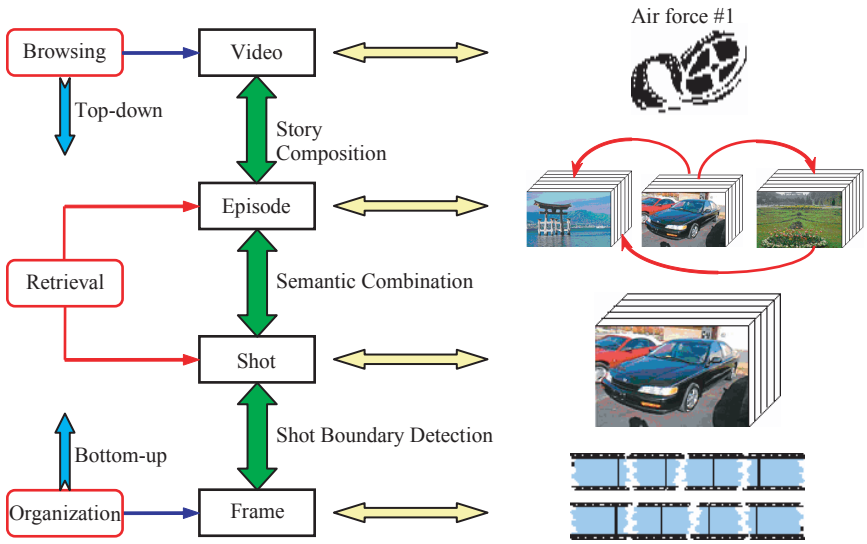
2001), feature extraction from the meaningful region layer to the visual perception layer, and object recognition from the visual perception layer to the object layer (Zhang, Gao, & Merzlyakov, 2002a). At the object layer, retrieval also can be performed with the help of self-adaptive relevance feedback (Gao, Zhang, & Yu, 2001). Such a progressive procedure could be considered as a synchronization of the procedure for progressive understanding of image contents. These various layers could provide distinct information of the image content, so this model is suitable to access from different levels. More details can be found in Zhang et al. (2004).

From the point of view of image understanding, the next stage would go beyond objects. The actions and interactions of objects and, thus, generated events (or scenes) are important in order to understand fully the contents of images. The images in this case would be described by some metadata.

Multi-Level Video Retrieval

Due to the great length and rich content of video data, quickly grasping a global picture or effectively retrieving pertinent information from such data becomes an challenging task. Organization and summarization of video data are the main approaches taken by content-based video retrieval.

Figure 5. Multi-level video organization



Video Organization

Video organization is a process of connecting and assembling the components of video in a predefined structure in order to provide a fast and flexible ability to browse and/or retrieve. In order to effectively organize the video for efficient browsing and querying, multi-level representation of video data is adopted. A typical hierarchical structure (scheme) for organization consists of four layers: video, episode, shot, and stream, as shown in Figure 5 (Zhang & Lu, 2002b). Three video operations—organization, browsing, and retrieval—are enabled by this scheme.

In contrast to normal browsing, which is a top-down process, organization is a bottom-up process. It starts from the lowest layer—the frame layer. This layer corresponds to the original video data that consist of a time sequence of frames. By shot boundary detection, the frames are grouped into shots. A shot is a basic unit of a video program. It consists of a number of frames that temporally are connected and spatially neighboring; it contains a continuous action in space. By using some high-level knowledge, several shots are combined to form an episode. Episode is a semantic unit that describes an act or a story. In other words, shots in an episode are content-related but can be separated temporally and/or disconnected spatially. A video program (e.g., movie) is built by a number of episodes that form a meaningful story. Retrieval can be conducted either in shot layer or episode layer by using various cues, such as global camera motion (frequently conveys the semantic implication of the video creator) and object motion vector (often represents intended action), as indicated by Yu and Zhang (2001a, 2001b).

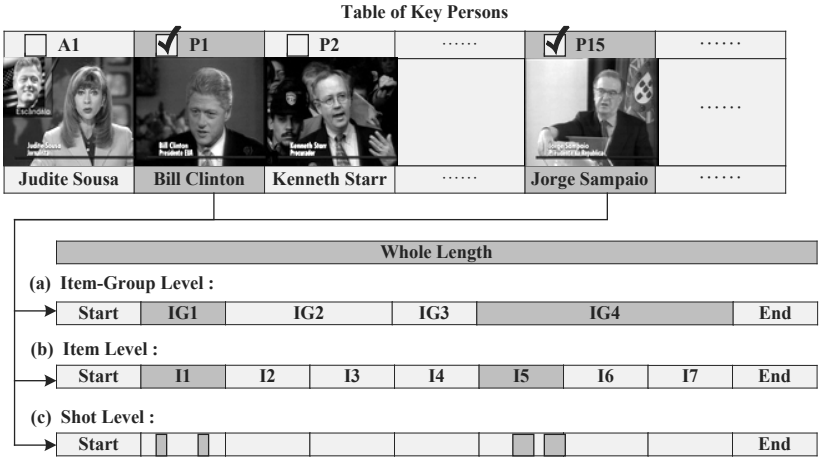
Video Summarization

Numerous techniques used in video organization (shot detection, shot clustering) also find their applications in video summarization, which is aimed at providing an abstraction and capturing essential subject matter with a compact representation of complex information. In addition, the object-based approach also could be employed. It first imposes spatial-temporal segmentation to get individual regions in video frames. Low-level features then are extracted for analysis from these objects instead of from frames directly. Therefore, semantic information can be expressed explicitly via the object features (Jiang & Zhang, 2005a).

With the help of these techniques, video summarization can be achieved. One framework for hierarchical abstraction of news programs with emphases on key persons (i.e., Main Speaker Close-Up, MSC, denoting camera-focused talking heads in center of the screen) has been proposed (Jiang & Zhang, 2005b). It can answer user queries such as to find all the key persons in a program or to highlight specific scenes with particular persons. The framework consists of three levels (from low physical level to high semantic level): shot level, item level, and item-group level. One actual example is shown in Figure 6 to provide an explanation.

Suppose a user is interested in two persons appearing in the news program (e.g., Bill Clinton and Jorge Sampaio). The user can find both of them from the previously constituted human table by name or by representative images and can check the corresponding boxes. Then, multi-level abstractions can be acquired: abstraction at shot level contains all the MSC shots of these two persons; abstraction at item level is made up of news items, including those MSC shots; abstraction at item-group level extends to every news item group with news items

Figure 6. Multi-level video abstraction



contained at item level. To obtain a summarization of preview sequence, frames belonging to all the video segments (at shot/item/item-group level) are concatenated together. Finally, abstractions at three levels with different durations are obtained in order to adaptively fulfill users' needs, from independent close-ups to complete video clips.

Overview of Current Focus

Current research works on SBVIR could be classified in several directions. A brief overview of some tracks is made in the following.

Image and Video Annotation

Image classification has been used to provide access to large image collections in a more efficient manner because the classification can reduce search space by filtering out the images in an unrelated category (Hirata et al., 2000). Image and video annotation significantly can enhance the performance of content-based visual retrieval systems by filtering out images from irrelevant classes during a search so as to reduce the processing time and the probability of error matching. One typical example is SIMPLIcity system, which uses integrated region matching based on image segmentation (Wang, Li, & Wiederhold, 2001). This system classifies images into various semantic categories such as textured/non-textured graph photograph. A series of statistical image classification methods also is designed during the development of this system for the purpose of searching images.

Inspired from SIMPLIcity system, a novel strategy by using feature selection in learning semantic concepts of image categories has been proposed (Xu & Zhang, 2006). For every image category, the salient patches on each image are detected by scale invariant feature transform (SIFT) introduced by Lowe (2004), and the 10-dimensional feature vectors are formed by PCA. Then, the DENCLUE (DENsity-based CLUstEring) clustering algorithm (Hinneburg & Keim, 1998) is applied to construct the continuous visual keyword dictionary.

A semantic description tool of multimedia content constructed with the Structured-Annotation Basic Tool of MPEG-7 multimedia description schemes (MDS) is presented in Kuo, Aoki, and Yasuda (2004). This tool annotates multimedia data with 12 main attributes regarding its semantic representation. The 12 attributes include answers to who, what, when, where, why, and how (5W1H) the digital content was produced, as well as the respective direction, distance, and duration (3D) information. With the proposed semantic attributes, digital multimedia contents are embedded as dozen dimensional digital content (DDDC). The establishment of DDDC would provide an interoperable methodology for multimedia content management applications at the semantic level.

A new algorithm for the automatic recognition of object classes from images (categorization) has been proposed (Winn, Criminisi, & Minka, 2005). It classifies a region according to the proportions of various visual words (clusters in feature space) by taking into account all pixels in images. The specific visual words and the characteristic proportions for each object are learned automatically by a supervised algorithm from a segmented training set, which gives the compact and yet discriminative appearance-based models for the object class.

According to a hypothesis that images dropped into the same text-cluster can be described with common visual features of those images, a strategy is described that combines textual and visual clustering results to retrieve images using semantic keywords and auto-annotate images based on similarity with existing keywords (Celebi & Alpkocak, 2005). Images first are clustered according to their text annotations using the C^3M clustering technique. The images also are segmented into regions and then clustered based on low-level visual features using the *k*-means clustering algorithm. The feature vector of the images then is mapped to a dimension equal to the number of visual clusters in which each entry of the new feature vector signifies the contribution of the image to that visual cluster. A feature vector also is created for the query image. Images in the textual cluster that give the highest matching score with the query image are determined by the matching of feature vectors.

A mechanism called Weblog-Style Video Annotation and Syndication to distribute and advertise video contents and to form communities centered on those contents effectively on the Internet is described in Yamamoto, Ohira, and Nagao (2005). This mechanism can tightly connect video contents, Weblogs, and users. It can be used for applications based on annotation by extracting the video annotations from the contents. In contrast to the machine processing of both voice- and image-recognition technology, which is difficult to apply to general contents, this mechanism has a low cost for annotation, as a variety of annotations is acquired from many users.

Human-Computer Interaction

In order to alleviate the problems that arise because of user subjectivity and the semantic gap, interactive retrieval systems have been proposed that place the user in the loop during retrievals. Such relevance feedback (RF) approaches aim to learn intended high-level query concepts and adjust for subjectivity in judgment by exploiting user input on successive iterations. In fact, since humans are much better than computers at extracting semantic information from images, relevance feedback has proved to be an effective tool for taking the user's judgment into account (Zhou & Huang, 2003). Generally, the user provides some quality assessment of the retrieval results to the system by indicating the degree of satisfaction with each of the retrieved results. The system then uses this feedback to adjust its query and/or the similarity measure in order to improve the next set of results.

Image retrieval is a complex processing task. In order to simplify the procedure, most current approaches for interactive retrieval make several restrictive assumptions (Kushki et al., 2004). One is that images considered similar according to some high-level concepts also fall close to each other in the low-level feature space. This generally is not true, as high-level semantic concepts may not be mapped directly to elements in a low-level feature space. Another is that the ideal conceptual query in the mind of a user can be represented in the low-level space and used to determine the region of this space that corresponds to images relevant to the query. However, as the mapping between low-level features and the conceptual space often is unclear, it is not possible to represent a high-level query as a single point in the low-level feature space. A novel approach has been proposed for interactive content-based image retrieval that provides user-centered image retrieval by lifting the previously mentioned restrictive assumptions imposed on existing CBIR systems while maintaining accurate retrieval performance (Kushki et al., 2004). This approach exploits user feedback

to generate multiple query images for similarity calculations. The final similarity between a given image and a high-level user concept then is obtained as a fusion of similarity of these images to a set of low-level query representations.

A new RF framework based on a feature selection algorithm that nicely combines the advantages of a probabilistic formulation with those of discriminative learning methods, using both the positive example (PE) and the negative example (NE), has been proposed (Kherfi & Ziou, 2006). It tries through interaction with the user to learn the weights the user assigns to image features according to their importance and then to apply the results obtained to define similarity measures that correspond better to the user's judgment for retrieval.

Models and Tools for Semantic Retrieval

Many mathematic models have been developed, and many useful tools from other fields have been applied for semantic retrieval, such as machine learning (Dai & Zhang, 2004; Lim & Jin, 2005). One powerful tool for such a work is data mining, especially for Web mining. The huge amounts of multivariate information offered by the Web have opened up new possibilities for many areas of research. Web mining refers to the use of data mining techniques to automatically retrieve, extract, and evaluate (generalize/analyze) information for knowledge discovery from Web documents and services (Arotaritei & Mitra, 2004). Web data typically are unlabeled, distributed, heterogeneous, semi-structured, time varying, and high-dimensional. Hence, any human interface needs to handle context-sensitive and imprecise queries and provide for summarization, deduction, personalization, and learning. A survey on Web mining involving fuzzy sets and their hybridization with other soft computing tools is presented in Arotaritei and Mitra (2004). The Web-mining taxonomy has been described. The individual functions like Web clustering, association rule mining, Web navigation, Web personalization, and Semantic Web have been discussed in the fuzzy framework.

Pattern recognition plays a significant role in content-based recognition. A list of pattern recognition methods developed for providing content-based access to visual information (both image and video) has been reviewed (Antani, Kasturi, & Jain, 2002). Here, the term *pattern recognition methods* refer to their applicability in feature extraction, feature clustering, generation of database indices, and determining similarity between the contents of the query and database elements.

A semantic learning method for content-based image retrieval using the analytic hierarchical process (AHP) has been proposed (Cheng et al., 2005). The AHP provides a good way to evaluate the fitness of a semantic description that is used to represent an image object. The idea behind this work is that the problem of assigning semantic descriptions to the objects of an image can be formulated as a multi-criteria preference problem, while AHP is a powerful tool for solving multi-criteria preference problems. In this approach, a semantic vector consisting of the values of fitness of semantics of a given image is used to represent the semantic content of the image according to a predefined concept hierarchy, and a method for ranking retrieved images according to their similarity measurements is made by integrating the high-level semantic distance and the low-level feature distance.

A new technique, cluster-based retrieval of images by unsupervised learning (CLUE), which exploits similarities among database images, for improving user interaction with image

retrieval systems has been proposed (Chen, Wang, & Krovetz, 2005). The major difference between a cluster-based image retrieval system and traditional CBIR systems lies in the two processing stages: selecting neighboring target images and image clustering, which are the major components of CLUE. CLUE retrieves image clusters by applying a graph-theoretic clustering algorithm to a collection of images in the vicinity of the query. In CLUE, the clustering process is dynamic, and the clusters formed depend on which images are retrieved in response to the query. CLUE can be combined with any real-valued symmetric similarity measure (metric or nonmetric). Thus, it may be embedded in many current CBIR systems, including relevance feedback systems.

A retrieval approach that uses concept languages to deal with nonverbally expressed information in multimedia is described in Lay and Gua (2006). The notion of concept language here is a rather loose one. It covers all other conventional methods of communication, including the systems of signs and rules, such as body language, painterly language, and the artificial languages of chess and the solitaire game. In operation, a finite number of elemental concepts of a concept language is identified and used to index multimedia documents. The elemental concepts then allow a large number of compound semantic queries to be expressed and operated as sentences of elemental concepts. Managing semantics by concept languages not only extends an intuitive query regime in which semantic queries can be specified more expressively and extensively but also allows concept detection to be restricted to a more manageable sum of semantic classes.

Miscellaneous Techniques in Application

Still other techniques could have potential in semantic-based visual information retrieval. A cascading framework for combining intra-image and interclass similarities in image retrieval motivated from probabilistic Bayesian principles has been proposed (Lim & Jin, 2005). Support vector machines are employed to learn local view-based semantics based on just-in-time fusion of color and texture features. A new detection-driven, block-based segmentation algorithm is designed to extract semantic features from images. The detection-based indexes also serve as input for support vector learning of image classifiers in order to generate class-relative indexes. During image retrieval, both intra-image and interclass similarities are combined to rank images. Such an approach would be suitable for unconstrained consumer photo images in which the objects are often ill-posed, occluded, and cluttered with poor lighting, focus, and exposure.

A novel, contents-based video retrieval system called DEV (Discussion Embedded Video), which combines video and an electronic bulletin board system (BBS), has been proposed (Haga & Kaneda, 2005). A BBS is used as an index of video data in the DEV system. The comments written in the BBS are arranged not according to the time that comments were submitted but by the playing time of video footage whose contents correspond to the contents of submitted comments. The main ideas behind this work are twofold: (1) since a participant's comments would be a summarization of the contents of a part of video, so it can be used as an index of it; and (2) as a user of this system can search part of the video by retrieving it via keywords in the BBS comments, so by detecting the part on which the comments from participants are concentrating it is possible to ascertain the topic of the video.

A two-stage retrieval process is described in Vogel and Schiele (2006). Users perform querying through image description by using a set of local semantic concepts and the size of the image area to be covered by the particular concept. In first stage, only small patches of the image are analyzed, whereas in the second stage, the patch information is processed, and the relevant images are retrieved. In this two-stage retrieval system, the precision and recall of retrieval can be modeled statistically. Based on the model, closed-form expressions that allow for the prediction as well as the optimization of the retrieval performance are designed.

Future Trends

There are several promising directions for future research.

Domain Knowledge and Learning

A semantic multimedia retrieval system consists of two components (Naphade & Huang, 2002). The first component links low-level physical attributes of multimedia data to high-level semantic class labels. The second component is domain knowledge or any such information apart from the semantic labels themselves that makes the system more competent to handle the semantics of the query. Many research works related to the first component have been conducted. For the second component, which can be in the form of rules, heuristics, or constraints, much more effort is required to automatically capture long-term human experience of human experts.

Various learning techniques such as active learning and multiple-instance learning are efficient to alleviate the cost of annotation (Naphade & Huang, 2002). One important direction is related to the development of an intelligent dialogue mechanism to increase the effectiveness of the user's feedback. The challenge is to attain performance that is considered useful by the end users of the systems. To this end, an active role of statistical learning in medium analysis will bridge the gap between the user's desire and the system's reply and will prevail in future media applications that involve semantic understanding.

What can be learned from human beings is not only domain knowledge but also the comportment of human beings. A technique employing the concept of small-world theory, which mimics the way in which humans keep track of descriptions of their friends and acquaintances, has been proposed (Androutsos, Androutsos, & Venetsanopoulos, 2006). Such a social networking behavior is extremely general to be applied to both low-level and semantic descriptors. This mirroring of the characteristics of social acquaintance provides an intelligent way to incorporate human knowledge into the retrieval process. Extending this procedure to other kinds of knowledge would be promoting.

Relevance Feedback and Association Feedback

Domain knowledge could be used in combination with relevance feedback to approach the optimal retrieval results. However, feedback is just a method for refining the results so it cannot totally determine the performance of retrieval systems. How to make this refining process fast following the user's aspiration in the course of retrieval is interesting (Zhang et al., 2004). When the system is based on lower-level features, relevance feedback only could improve the retrieval results to some degree. The use of relevance feedback based on high-level content description in the object level could further improve the performance.

On the other side, a potential direction would be to use association feedback based on feature elements (Xu & Zhang, 2001, 2002). Association feedback tries to find out the associated parts between the existing interest (user intent) and the new target (related to the current retrieval results), which usually is a subset of the demand feature element set; that is, the bridge to the new retrieval.

Even Higher-Level Exploration

Semantic retrieval requires the use of a cognitive model—a feature element construction model that tries to enhance the view-based model—while importing some useful inferences from image-based theory has been proposed (Xu & Zhang, 2003). A feature element is the discrete unit extracted from low-level data that represent the distinct or discriminating visual characteristics that may be related to the essence of the objects.

The semantic level is higher than the feature level, while the affective level is higher than the semantic level (Hanjalic, 2001). Affection is associated with some abstract attributes that are quite subjective. For example, atmosphere is an important abstract attribute for film frames. Atmosphere serves an important role in generating the scene's topic or in conveying the message behind the scene's story. Five typical categories of atmosphere semantics have been studied: vigor and strength, mystery or ghastfulness, victory and brightness, peace or desolation, and lack unity and appears disjoint (Xu & Zhang, 2005).

Conclusion

There is no clear image yet about the development of semantic-based visual information retrieval. Research made from a low feature level to a high semantic level in visual information retrieval already has obtained the result of numerous noteworthy progresses and quite a number of publications. However, many approaches have obtained only limited success, and various investigations are still pursued on the way. For example, still very little work has been done to address the issue of human perception of visual data content (Vogel & Schiele, 2006). Several practical approaches, such as multi-level model, classification and annotation, machine-learning techniques, human-computer interaction, as well as various models and tools, have been discussed in this chapter. Few potential research directions,

such as the domain knowledge and learning, relevance feedback and association feedback, as well as research at an even high level, are discussed. It is apparent that the future for content-based visual information retrieval will rely on high-level research.

Acknowledgments

This work has been supported by Grants NNSF-60573148 and SRFDP-20050003013.

References

Amir, A., & Lindenbaum, M. (1998). A generic grouping algorithm and its quantitative analysis. *IEEE PAMI, 20*(2), 168–185.

Androutsos, P., Androutsos, D., & Venetsanopoulos, A. N. (2006). Small world distributed access of multimedia data. *IEEE Signal Processing Magazine, 23*(2), 142–153.

Antani, S., Kasturi, R., & Jain, R. (2002). A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern Recognition, 35*(4), 945–965.

Arotaritei, D., & Mitra, S. (2004). Web mining: A survey in the fuzzy framework. *Fuzzy Sets and Systems, 148*, 5–19.

Bimbo, A. (1999). *Visual information retrieval*. Morgan Kaufmann.

Castelli, V., Bergman, L. D., Kontoyiannis, I., et al. (1998). Progressive search and retrieval in large image archives. *IBM J. Res. Develop., 42*(2), 253–268.

Celebi, E., & Alpkocak, A. (2005). Combining textual and visual clusters for semantic image retrieval and auto-annotation. In *Proceedings of the 2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology* (pp. 219–225).

Chang, S. F., Chen, W., Meng, H. J., Sundaram, H., & Zhong, D. (1998). A fully automated content-based video search engine supporting spatiotemporal queries. *IEEE CSVT, 8*(5), 602–615.

Chen, Y. X., Wang, J. Z., & Krovetz, R. (2005). CLUE: Cluster-based retrieval of images by unsupervised learning. *IEEE IP, 14*(8), 1187–1201.

Cheng, S. C., Chou, T. C., Yang, C. L., et al. (2005). A semantic learning for content-based image retrieval using analytical hierarchy process. *Expert Systems with Applications, 28*(3), 495–505.

Dai, S. Y., & Zhang, Y. J. (2004). Adaboost in region-based image retrieval. In *Proceedings of the International Conference on Acoustic, Speech, and Signal Processing* (pp. 429–432).

Dai, S. Y., & Zhang, Y. J. (2005). Unbalanced region matching based on two-level description for image retrieval. *Pattern Recognition Letters*, *26*(5), 565–580.

Doermann, D. (1998). The indexing and retrieval of document images: A survey. *Computer Vision and Image Understanding, 70*(3), 287–298.

Gao, Y. Y., Zhang, Y. J., & Merzlyakov, N. S. (2000). Semantic-based image description model and its implementation for image retrieval. In *Proceedings of the First International Conference on Image and Graphics* (pp. 657–660).

Gao, Y. Y., Zhang, Y. J., & Yu, F. (2001). Self-adaptive relevance feedback based on multilevel image content analysis. *SPIE*, *4315*, 449–459.

Haga, H., & Kaneda, H. (2005). A usability survey of a contents-based video retrieval system by combining digital video and an electronic bulletin board. *Internet and Higher Education, 8*(3), 251–262.

Hanjalic, A. (2001). Video and image retrieval beyond the cognitive level: The needs and possibilities. *SPIE, 4315*, 130–140.

Hinneburg, A., & Keim, D. A. (1998). An efficient approach to clustering in large multimedia databases with noise. *KDD'98*, 58–65.

Hirata, K., et al. (2000). Integration of image matching and classification for multimedia navigation. *Multimedia Tools and Applications, 11*(3), 295–309.

Hong, D. Z., Wu, J. K., & Singh, S. S. (1999). Refining image retrieval based on context-driven method. *SPIE, 3656*, 581–593.

Jaimes, A., & Chang, S. F. (1999). Model-based classification of visual information for content-based retrieval. *SPIE, 3656*, 402–414.

Jiang, F., & Zhang, Y. J. (2005a). Camera attention weighted strategy for video shot grouping. *SPIE, 5960*, 428–436.

Jiang, F., & Zhang, Y. J. (2005b). News video indexing and abstraction by specific visual cues: MSC and news caption. In S. Deb (Ed.), *Video data management and information retrieval* (pp. 254–281). IRM Press.

Kato, T. (1992). Database architecture for content-based image retrieval. *SPIE, 1662*, 112–123.

Kherfi, M. L., & Ziou, D. (2006). Relevance feedback for CBIR: A new approach based on probabilistic feature weighting with positive and negative examples. *IEEE IP, 15*(4), 1014–1033.

Kuo, P. J., Aoki, T., & Yasuda, H. (2004). MPEG-7 based dozen dimensional digital content architecture for semantic image retrieval services. In *Proceedings of the 2004 IEEE International Conference on E-Technology, E-Commerce and E-Service (EEE'04)* (pp. 517–524).

Kushki, A., et al. (2004). Query feedback for interactive image retrieval. *IEEE CSVT, 14*(5), 644–655.

Lay, J. A., & Gua, L. (2006). Semantic retrieval of multimedia by concept languages. *IEEE Signal Processing Magazine, 23*(2), 115–123.

Lim, J. H., & Jin, J. S. (2005). Combining intra-image and inter-class semantics for consumer image retrieval. *Pattern Recognition, 38*(6), 847–864.

Lowe, D. G. (2004). Distinctive image features from scale-invariant key-points. *International Journal of Computer Vision, 60*(2), 91–110.

Luo, J. B., Boutell, M., & Brown, C. (2006). Pictures are not taken in a vacuum: An overview of exploiting context for semantic scene content understanding. *IEEE Signal Processing Magazine, 23*(2), 101–114.

Luo, Y., et al. (2001). Extracting meaningful region for content-based retrieval of image and video. *SPIE*, *4310*, 455–464

Naphade, M. R., & Huang, T. S. (2002). Extracting semantics from audiovisual content: The final frontier in multimedia retrieval. *IEEE NN, 13*(4), 793–810.

Rui, Y., Huang, T. S., & Chang, S.F. (1999). Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation, 10*(1), 39–62.

Smeulders, A. W. M., et al. (2000). Content-based image retrieval at the end of the early years. *IEEE PAMI, 22*(12), 1349–1380.

Vogel, J., & Schiele, B. (2006). Performance evaluation and optimization for content-based image retrieval. *Pattern Recognition, 39*, 897–909.

Wang, J. Z., Li, J., & Wiederhold, G. (2001). SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE PAMI, 23*(9), 947–963.

Winn, J., Criminisi, A., & Minka, T. (2005). Object categorization by learned universal visual dictionary. In *Proceedings of the 10th International Conference on Computer Vision (ICCV'05)* (Vol. 2, pp. 1800–1807).

Xu, Y., & Zhang, Y. J. (2001). Association feedback: A novel tool for feature elements based image retrieval. In *Proceedings of the Second IEEE Pacific Rim Conference on Multimedia* (pp. 506–513).

Xu, Y., & Zhang, Y. J. (2002). Feature element theory for image recognition and retrieval. *SPIE*, *4676*, 126–137

Xu, Y., & Zhang, Y. J. (2003). Semantic retrieval based on feature element constructional model and bias competition mechanism. *SPIE*, *5021*, 77–88.

Xu, F., & Zhang, Y. J. (2005). Atmosphere-based image classification through illumination and hue. *SPIE, 5960*, 596–603.

Xu, F., & Zhang, Y. J. (2006). Feature selection for image categorization. In *Proceedings of the Seventh Asian Conference on Computer Vision* (pp. 653–662).

Yamamoto, D., Ohira, S., & Nagao, K. (2005). Weblog-style video annotation and syndication. In *Proceedings of the First International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS'05)* (pp. 1–4).

Yu, T. L., & Zhang, Y. J. (2001a). Motion feature extraction for content-based video sequence retrieval. *SPIE, 4311*, 378–388.

Yu, T. L., & Zhang, Y. J. (2001b). Retrieval of video clips using global motion information. *IEE Electronics Letters, 37*(14), 893–895.

Zhang, Y. J. (2003). *Content-based visual information retrieval*. Beijing, China: Science Publisher.

Zhang, Y. J. (2005a). Advanced techniques for object-based image retrieval. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (Vol. 1, pp. 68–73). Hershey, PA: Idea Group Reference.

Zhang, Y. J. (2005b). New advancements in image segmentation for CBIR. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (Vol. 4, pp. 2105–2109). Hershey, PA: Idea Group Reference.

Zhang, Y. J. (2006). Mining for image classification based on feature elements. In J. Wang (Ed.), *Encyclopedia of Data Warehousing and Mining* (Vol. 1, pp. 773–778).

Zhang, Y. J., Gao, Y. Y., & Luo, Y. (2004). Object-based techniques for image retrieval. In S. Deb (Ed.), *Multimedia systems and content-based image retrieval* (pp. 156–181). Hershey, PA: Idea Group Publishing.

Zhang, Y. J., Gao, Y. Y., & Merzlyakov, N. S. (2002a). Object recognition and matching for image retrieval. *SPIE, 4875*, 1083–1089.

Zhang, Y. J., & Lu, H. B. (2002b). A hierarchical organization scheme for video data. *Pattern Recognition*, *35*(11), 2381–2387.

Zhou, X. S., & Huang, T. S. (2003). Relevance feedback for image retrieval: A comprehensive review. *Multimedia Systems, 8*(6), 536–544.