Content-Based Multimedia Information Retrieval: State of the Art and Challenges

MICHAEL S. LEW
Leiden University, The Netherlands
NICU SEBE
University of Amsterdam, The Netherlands
CHABANE DJERABA
LIFL, France
and
RAMESH JAIN
University of California at Irvine, USA

Extending beyond the boundaries of science, art, and culture, content-based multimedia information retrieval provides new paradigms and methods for searching through the myriad variety of media all over the world. This survey reviews 100+ recent articles on content-based multimedia information retrieval and discusses their role in current research directions which include browsing and search paradigms, user studies, affective computing, learning, semantic queries, new features and media types, high performance indexing, and evaluation techniques. Based on the current state of the art, we discuss the major challenges for the future.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning; I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms: Design, Experimentation, Human Factors, Performance

Additional Key Words and Phrases: Multimedia information retrieval, image search, video retrieval, audio retrieval, image databases, multimedia indexing, human-computer interaction

1. INTRODUCTION

Multimedia information retrieval (MIR) is about the search for knowledge in all its forms, everywhere. Indeed, what good is all the knowledge in the world if it is not possible to find anything? This sentiment is mirrored as an ACM SIGMM grand challenge [Rowe and Jain 2005]: "make capturing, storing, finding, and using digital media an everyday occurrence in our computing environment."

This article is meant for researchers in the area of content-based retrieval of multimedia. Currently, the fundamental problem has been how to enable or improve multimedia retrieval using content-based methods. Content-based methods are necessary when text annotations are nonexistent or incomplete. Furthermore, content-based methods can potentially improve retrieval accuracy even when text annotations are present by giving additional insight into the media collections.

Authors' address: N. Sebe, University of Amsterdam, Faculty of Science, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands.

Our search for digital knowledge began several decades ago when the idea of digitizing media was commonplace but when books were still the primary medium for storing knowledge. Before the field of multimedia information retrieval coalesced into a scientific community, there were many contributory advances from a wide set of established scientific fields. From a theoretical perspective, areas such as artificial intelligence, optimization theory, computational vision, and pattern recognition contributed significantly to the underlying mathematical foundation of MIR. Psychology and related areas such as aesthetics and ergonomics provided basic foundations for the interaction with the user. Furthermore, applications of pictorial search into a database of imagery already existed in niche forms such as face recognition, robotic guidance, and character recognition.

The earliest years of MIR were frequently based on computer vision (three excellent books, Ballard and Brown [1982]; Levine [1985]; and Haralick and Shapiro 1993]) algorithms which focused on feature-based similarity search over images, video, and audio. Influential and popular examples of these systems are QBIC [Flickner et al. 1995] and Virage [Bach et al. 1996], circa mid 90s. Within a few years, the basic concept of the similarity search was transferred to several Internet image search engines including Webseek [Smith and Chang 1997] and Webseer [Frankel et al. 1996]. Significant effort was also placed on the direct integration of the feature-based similarity search into enterprises-level databases such as Informix datablades, IBM DB2 Extenders, or Oracle Cartridges [Bliujute et al. 1999; Egas et al. 1999] to make MIR more accessible to private industry.

In the area of video retrieval, the main focus in the mid 90s was on robust shot boundary detection; the most common approaches involved thresholding the distance between color histograms corresponding to two consecutive frames in a video [Flickner et al. 1995]. Hanjalic et al. [1997] proposed a method which overcame the problem of subjective user thresholds. Their approach was not dependent on any manual parameters. It gave a set of keyframes based on an objective model for the video information flow. Haas et al. [1997] described a method of using the motion within the video to determine the shot boundary locations. Their method outperformed the histogram approaches of the period and also performed semantic classification of the video shots into categories such as zoom-in, zoom-out, pan, and so on. A more recent practitioner's guide to video transition detection is given by Lienhart [2001].

Near the turn of the 21st century, researchers noticed that the feature-based similarity search algorithms were not as intuitive or user-friendly as they had expected. One could say that systems built by research scientists were essentially systems which could only be used effectively by scientists. The new direction was geared toward designing systems which would be user-friendly and could bring the vast multimedia knowledge from libraries, databases, and collections to the world. To do this, it was noted that the next evolution of systems would need to understand the semantics of a query, not simply the low-level underlying computational features. This general problem was called "bridging the semantic gap". From a pattern recognition perspective, this roughly meant translating the easily computable low-level content-based media features to high-level concepts or terms which would be intuitive to the user. Examples of bridging the semantic gap for the single concept of human faces were demonstrated by Rowley et al. [1996] and Lew and Huijsmans [1996]. Perhaps the earliest pictorial content-based retrieval system which addressed the semantic gap problem in the query interface, indexing, and results was the ImageScape search engine [Lew 2000]. In this system, the user could make direct queries for multiple visual objects such as sky, trees, water, and so on, using spatially positioned icons in a WWW index containing 10+ million images and videos using keyframes. The system used information theory to determine the best features for minimizing uncertainty in the classification.

At this point, it is important to note that the feature-based similarity search engines were useful in a variety of contexts [Smeulders et al. 2000] such as searching trademark databases [Eakins et al. 2003], finding video shots with similar visual content and motion, or for DJs searching for music with similar

rhythms [Foote 1999], and automatic detection of pornographic content [Forsyth and Fleck 1999; Bosson et al. 2002]. Intuitively, the most pertinent applications are those where the basic features such as color and texture in images and video, or dominant rhythm, melody, or frequency spectrum in audio [Foote 1999] are highly correlated to the search goals of the particular application.

2. RECENT WORK

In this section, we discuss representative work [Dimitrova 2003; Lew 2001; Sebe et al. 2003b] done in content-based multimedia retrieval in recent years. Sections 3 and 4 discuss and directly address important challenges for the future. The two fundamental necessities for a multimedia information retrieval system are (1) searching for a particular media item, and (2) browsing and summarizing a media collection. In searching for a particular media item, the current systems have significant limitations, such as an inability to understand a wide user vocabulary and the user's satisfaction level, nor do there exist credible representative real-world test sets for evaluation or even benchmarking measures which are clearly correlated with user satisfaction. In general, current systems have not yet had significant impact on society due to an inability to bridge the semantic gap between computers and humans.

The prevalent research topics which have potential for improving multimedia retrieval by bridging the semantic gap are as follows: human-centered computing, new features, new media, browsing and summarization, and evaluation/benchmarking. In human-centered computing, the main idea is to satisfy the user and allow the user to make queries in their own terminology. User studies give us insight directly into the interactions between human and computer. Experiential computing also focuses on methods for allowing the user to explore and gain insights into media collections. On a fundamental level, the notion of user satisfaction is inherently emotional. Affective computing is fascinating because it focuses on understanding the user's emotional state and intelligently reacting to it. It can also be beneficial in measuring user satisfaction in the retrieval process.

Learning algorithms are interesting because they potentially allow the computer to understand the media collection on a semantic level. Furthermore, learning algorithms may be able to adapt and compensate for the noise and clutter in real-world contexts. New features are pertinent in that they can potentially improve the detection and recognition process or be correlated with human perception. New media types address the changing nature of the media in the collections or databases. Some of the recent new media include 3D models (i.e., for virtual reality or games) and biological imaging data (i.e., for understanding the machinery of life). As scientists, we need to objectively evaluate and benchmark the performance of the systems and take into account factors such as user satisfaction with results. Currently, there are no large international test sets for the many problems such as searching personal media collections so significant effort has been focused on developing paradigms which are effective for evaluation. Furthermore, as collections grow from gigabyte to terabyte to petabyte sizes, high performance algorithms will be necessary in order to respond to a query in an acceptable time period.

Currently, the most commonly used test sets include collections involving personal photos, web images and videos, cultural heritage images, news video, and the Corel stock photography collection which is also the most frequently mentioned collection. We are not asserting that the Corel collection is a good test set. We suspect it is popular simply because it is widely available and related loosely to real-world usage. Furthermore, we think that it is only representative and suitable if the main goal of the particular retrieval system is to find professional stock photography.

For the most recent research, there currently are several conferences dedicated to the field of MIR such as the ACM SIGMM Workshop on Multimedia Information Retrieval (http://www.liacs.nl/~mir) and the International Conference on Image and Video Retrieval (http://www.civr.org). For a searchable

MIR library, we suggest the community driven digital library at the Association for Multimedia Search and Retrieval (http://www.amsr.org). Additionally, the general multimedia conferences such as ACM Multimedia (http://www.sigmm.org) and the IEEE International Conference on Multimedia and Expo (ICME) typically have MIR-related tracks.

2.1 Human-Centered

By human-centered we mean systems which consider the behavior and needs of the human user [Jaimes and Sebe 2006]. As noted earlier, the foundational areas of MIR were often in computing-centric fields. However, since the primary goal is to provide effective browsing and search tools for the user, it is clear that the design of the systems should be human-centric. There have been several major recent initiatives in this direction such as user understanding, experiential computing, and affective computing.

One of the most fascinating studies was done on whether organization by similarity assists image browsing [Rodden 2001]. The users were asked to illustrate a set of destination guide articles for a travel Web site. The similarity by visual content view was compared with a text caption similarity view. In 40 of the 54 searches, users chose to use the text caption view with comments such as "it gave me a breakdown of the subject." In many cases, the users began with the text caption view to ensure sufficient diversity. Also, it was noted by the users that they would want both possibilities simultaneously. In another experiment, the visual similarity view was compared with a random set. Most users were slightly more satisfied with the visual similarity view, but there was one user who preferred the random images view. Specifically, the visual similarity view was preferred in 66% of the searches.

A good description of user requirements for photoware is discussed in Frohlich [2002] and Lim et al. [2003]. The importance of time in user interfaces is discussed in Graham et al. [2002]. By understanding user types [Enser and Sandom 2003; Rubin 2004; Enser et al. 2005], it is clear that the current work has not addressed the full plurality of image and user types and that a broad evaluation is important. In specific cases, there has been niche work such as the use of general purpose documentary images by generalist and specialist users [Markkula and Sormunen 2000] and the use of creative images by specialist users [Hastings 1999]. Other interesting studies have been done on the process of managing personal photograph collections [Rodden and Wood 2003]. Worring and Gevers [2001] describe a concise analysis of methodologies for interactive retrieval of color images which includes guidelines for selecting methods based on the domain and the type of search goal. Also, Worring et al. [2004] give useful insights into how users apply the steps of indexing, filtering, browsing, and ranking in video retrieval. Usage mining in large multimedia databases is another emerging problem. The objective is to extract the hidden information in user behaviors on large multimedia databases. A framework for video usage mining has been presented in Mongy et al. [2005].

The idea behind experiential computing [Jain 2003; Jain et al. 2003] is that decision makers routinely need insights that come purely from their own experience and experimentation with media and applications. These insights come from multiple perspectives and exploration [Gong et al. 2004]. Instead of analyzing an experience, experiential environments provide support for naturally understanding events. In the context of MIR, experiential environments provide interfaces for creatively exploring sets of data, giving multiple perspectives, and allowing the user to follow his insights.

Affective computing [Picard 2000; Berthouze and Kato 1998; Hanjalic and Xu 2005] seeks to provide better interaction with the user by understanding the user's emotional state and responding in a way which influences or takes into account the user's emotions. For example, Sebe et al. [2002] recognize emotions automatically using a Cauchy classifier on an interactive 3D wireframe model of the face. Wang et al. [2004] examine the problem of grouping images into emotional categories. They introduce a novel feature based on line direction-length which works effectively on a set of art paintings. Salway

and Graham [2003] develop a method for extracting character emotions from film which is based on a model that links character emotions to the events in their environment.

2.2 Learning and Semantics

The potential for learning in multimedia retrieval is quite compelling in bridging the semantic gap, and the recent research literature has seen significant interest in applying classification and learning [Therrien 1989; Winston 1992; Haralick and Shapiro 1993] algorithms to MIR. The Karhunen-Loeve (KL) transform or principal components method [Therrien 1989] has the property of representational optimality for a linear description of the media. It is important to distinguish between representational optimality versus classification optimality. The ability to optimally represent a class does not necessarily lead to optimally classifying an instance of the class. An example of an improvement on the principal component approach was proposed by Capelli et al. [2001] where they suggest a multispace KL for classification purposes. The multispace KL directly addresses the problem of when a class is represented by multiple clusters in feature space and can be used in most cases where the normal KL would be appropriate. Zhou and Huang [2001] compared discriminating transforms and SVM for image retrieval. They found that the biased discriminating transform (BDT) outperformed the SVM. Lew and Denteneer [2001] found that the optimal linear keys (in the sense of minimizing the distance between two relevant images) could be found directly from Fisher's Linear Discriminant. Liu et al. [2003] find optimal linear subspaces by formulating the retrieval problem as optimization on a Grassman manifold. Balakrishnan et al. [2005] propose a new representation based on biological vision which uses complementary subspaces. They compare their new representation with principal component analysis, the discrete cosine transform, and the independent component transform.

Another approach to learning semantics is to determine the associations behind features and the semantic descriptions. Djeraba [2002, 2003] examines the problem of data mining and discovering hidden associations during image indexing and considers a visual dictionary which groups together similar colors and textures. A learning approach is explored by Krishnapuram et al. [2004] in which they introduce a fuzzy graph matching algorithm. Greenspan et al. [2004] performs clustering on space-time regions in feature space in an effort to create a piece-wise GMM framework which allows for the detection of video events.

2.2.1 *Concept Detection in Complex Backgrounds* . One of the most important challenges and perhaps the most difficult problem in semantic understanding of media is visual concept detection in the *presence of complex backgrounds*. Many researchers have looked at classifying whole images, but the granularity is often too coarse to be useful in real-world applications. Its typically necessary to find the human in the picture, not simply global features. Another limiting case is when researchers examined the problem of detecting visual concepts in laboratory conditions where the background is simple and, therefore, can be easily segmented. Thus, the challenge is to detect all of the semantic content within an image such as faces, trees, animals, and so on, with emphasis on the presence of complex backgrounds.

In the mid 90s, there was a great deal of success in the special case of detecting the locations of human faces in grayscale images with complex backgrounds. Lew and Huijsmans [1996] used Shannon's information theory to minimize the uncertainty in the face detection process. Rowley et al. [1996] applied several neural networks to detect faces. Both methods had the limitation of searching for whole faces which prompted later component-based model approaches that combined separate detectors for the eyes and nose regions. In the case of near frontal face views in high-quality photographs, the early systems generally performed near 95% accuracy with minimal false positives. Nonfrontal views and low-quality or older images from cultural heritage collections are still considered to be very difficult.

An early example of designing a simple detector for city pictures was demonstrated by Vailaya et al. [1998]. They used a nearest neighbor classifier in conjunction with edge histograms. In more recent work, Schneiderman and Kanade [2004] proposed a system for component-based face detection using the statistics of parts. Chua et al. [2002] used the gradient energy directly from the video representation to detect faces based on the high contrast areas such as the eyes, nose, and mouth. They also compared a rules-based classifier with a neural network and found that the neural network gave superior accuracy. For a good overview, Yang et al. [2002] did a comprehensive survey on the area of face detection.

Detecting a wider set of concepts other than human faces turned out to be fairly difficult. In the context of image search over the Internet, Lew [2000] showed a system for detecting sky, trees, mountains, grass, and faces in images with complex backgrounds. Fan et al. [2004] used multilevel annotation of natural scenes utilizing dominant image components and semantic concepts. Li and Wang [2003] used a statistical modeling approach to convert images to keywords. Rautianinen et al. [2001] used temporal gradients and audio analysis in video to detect semantic concepts.

In certain contexts, there may be several media type available which allows for multimodal analysis. Shen et al. [2000] discussed a method for giving descriptions of WWW images by using lexical chain analysis of the nearby text on Web pages. Benitez and Chang [2002] exploit WordNet to disambiguate descriptive words. They also found a 3–15% improvement in combining pictorial search with text analysis. Amir et al. [2004] proposed a framework for a multimodal system for video event detection which combined speech recognition and annotated video. Dimitrova et al. [2000] proposed a Hidden Markov Model, using text and faces for video classification. In the TRECVID [Smeaton and Over 2003] project, the current focus is on multiple domain concept detection for video retrieval.

2.2.2 *Relevance Feedback.* Beyond the one-shot queries in the early similarity-based search systems, the next generation of systems attempted to integrate continuous feedback from the user in order to learn more about the user query. The interactive process of asking the user a sequential set of questions after each round of results was called *relevance feedback* because of its similarity to older pure text approaches. Relevance feedback can be considered a special case of *emergent semantics*. Other names have included query refinement, interactive search, and active learning from the computer vision literature.

The fundamental idea behind relevance feedback is to show the user a list of candidate images, ask the user to decide whether each image is relevant or irrelevant, and modify the parameter space, semantic space, feature space, or classification space to reflect the relevant and irrelevant examples. In the simplest relevance feedback method from Rocchio [1971], the idea is to move the query point toward the relevant examples and away from the irrelevant examples. In principle, one general view is to view relevance feedback as a particular type of pattern classification in which the positive and negative examples are found from the relevant and irrelevant labels, respectively.

Therefore, it is possible to apply any learning algorithm into the relevance feedback loop. One of the major problems in relevance feedback is how to address the small training set. A typical user may only want to label 50 images when the algorithm really needs 5000 examples. If we compare the simple Rocchio algorithm to more sophisticated learning algorithms such as neural networks, it is clear that one reason the Rocchio algorithm is popular is that it requires very few examples. However, one challenging limitation of the Rocchio algorithm is that there is a single query point which refers to a single cluster of results. In the discussion that follows, we briefly describe some of the recent innovations in relevance feedback.

Chang et al. [1998] proposed a framework which allows for interactive construction of a set of queries that detect visual concepts such as sunsets. Sclaroff et al. [2001] describe the first WWW image search

engine which focused on relevance feedback-based improvement of the results. In their initial system, where they used relevance feedback to guide the feature selection process, it was found that the positive examples were more important in maximizing accuracy than the negative examples. Rui and Huang [2001] compare heuristic to optimization-based parameter updating and find that the optimization-based method achieves higher accuracy.

Chen et al. [2001] described a one-class SVM method for updating the feedback space which shows substantially improved results over previous work. He et al. [2002] use both short-term and long-term perspectives to infer a semantic space from user's relevance feedback for image retrieval. The short-term perspective was found by marking the top 3 incorrect examples from the results as irrelevant and selecting at most 3 images as relevant examples from the current iteration. The long-term perspective was found by updating the semantic space from the results of the short term perspective. Yin et al. [2005] found that combining multiple relevance feedback strategies gives superior results as opposed to any single strategy. Tieu and Viola [2004] proposed a method for applying the AdaBoost learning algorithm and noted that it is quite suitable for relevance feedback due to the fact that AdaBoost works well with small training sets. Howe [2003] compares different strategies using AdaBoost. Dy et al. [2003] use a two-level approach via customized queries and introduce a new unsupervised learning method called feature subset selection using expectation-maximization clustering. Their method doubled the accuracy for the case of a set of lung images. Guo et al. [2001] performed a comparison between AdaBoost and SVM and found that SVM gives superior retrieval results. Haas et al. [2004] described a general paradigm which integrates external knowledge sources with a relevance feedback mechanism and demonstrated on real test sets that the external knowledge substantially improves the relevance of the results. A good overview can also be found from Muller et al. [2000].

2.3 New Features and Similarity Measures

Research did not only proceed along the lines of improved search algorithms, but also toward creating new features and similarity measures based on color, texture, and shape. One of the recent interesting additions to the set of features are from the MPEG-7 standard [Pereira and Koenen 2001]. The new color features [Lew 2001; Gevers 2001] such as the NF, rgb, and m color spaces have specific benefits in areas such as lighting invariance, intuitiveness, and perceptual uniformity. A quantitative comparison of influential color models is performed in Sebe and Lew [2001].

In texture understanding, Ojala et al. [1996] found that combining relatively simple texture histograms outperformed traditional texture models such as Gaussian or Markov features. Jafari-Khouzani and Soltanian-Zadeh [2005] proposed a new texture feature based on the Radon transform orientation which has the significant advantage of being rotationally invariant. Insight into the MPEG-7 texture descriptors is given by Wu et al. [2001].

Veltkamp and Hagedoorn [2001] describe the state-of-the-art in shape matching from the perspective of computational geometry. Sebe and Lew [2002] evaluate a wide set of shape measures in the context of image retrieval. Srivastava et al. [2005] describe some novel approaches to learning shape. Sebastian et al. [2004] introduce the notion of shape recognition using shock graphs. Bartolini et al. [2005] suggest using the Fourier phase and time warping distance.

Foote [2000] introduces a feature for audio based on local self-similarity. The important benefit of the feature is that it can be computed for any audio signal and works well on a wide variety of audio segmentation and retrieval applications. Bakker and Lew [2002] suggest several new audio features called the frequency spectrum differentials and the differential swap rate. They evaluate the new audio features in the context of automatic labeling of the sample as either speech, music, piano, organ, guitar, automobile, explosion, or silence and achieve promising results.

Equally important to novel features is the method to determine similarity between them. Jolion [2001] gives an excellent overview of the common similarity measures. Sebe et al. [2000] discuss how to derive an optimal similarity measure given a training set. In particular, they find that the sum of squared distance tends to be the worst similarity measure and that the Cauchy metric outperforms the commonly used distance measures. Jacobs et al. [2000] investigates nonmetric distances and evaluates their performance. Beretti et al. [2001] propose an algorithm which relies on graph matching for a similarity measure. Cooper et al. [2005] suggest measuring image similarity using time and pictorial content.

In the last decades, a lot of research has been done on the matching of images and their structures [Schmid et al. 2000; Mikolajczyk and Schmid 2004]. Although the approaches are very different, most methods use some kind of point selection from which descriptors are derived. Most of these approaches address the detection of points and regions that can be detected in an affine invariant way.

Lindeberg [1998] proposed an interesting scale-level detector which is based on determining maxima over scale of a normalized blob measure. The Laplacian-of-Gaussian (LoG) function is used for building the scale space. Mikolajczyk and Schmid [2004] showed that this function is very suitable for automatic scale selection of structures. An efficient algorithm to be used in object recognition was proposed by Lowe [2004]. This algorithm constructs a scale space pyramid using difference-of-Gaussian (doG) filters. The doG can be used to obtain an efficient approximation of the LoG. From the local 3D maxima, a robust descriptor is built for matching purposes. The disadvantage of using doG or LoG as feature detectors is that the repeatability is not optimal since they not only respond to blobs, but also to high gradients in one direction. Because of this, the localization of the features may not be very accurate.

An approach that intuitively arises from this observation is the separation of the feature detector and the scale selection. The commonly used Harris detector [Harris and Stephens 1988] is robust to noise and lighting variations, but only to a very limited extent to scale changes [Schmid et al. 2000]. To deal with this, Dufournoud et al. [2000] proposed the scale-adapted Harris operator. Given the scale-adapted Harris operator, a scale space can be created. Local 3D maxima in this scale space can be taken as salient points, but this scale-adapted Harris operator rarely attains a maximum over scales. This results in very few points which are not representative enough for the image. To address this problem, Mikolajczyk and Schmid [2004] proposed the Harris-Laplace detector that merges the scale-adapted Harris corner detector and the Laplacian-based scale selection.

During the last few years, much of the research on scale invariance has been generalized to affine invariance. Affine invariance is defined here as invariance to nonuniform scaling in different directions. This allows for matching of descriptors under perspective transformations since a global perspective transformation can be locally approximated by an affine transformation [Tuytelaars and van Gool 2000]. The use of the second moment matrix (or autocorrelation matrix) of a point for affine normalization was explored by Lindeberg and Garding [1997]. A similar approach was used by Baumberg [2000] for feature matching.

All the methods discussed were designed to be used in the context of object-class recognition application. However, it was found that wavelet-based salient points [Tian et al. 2001] outperform traditional interest operators such as corner detectors when they are applied to general content-based image retrieval. For a good overview, we refer the reader to Sebe et al. [2003a].

2.4 New Media

In the early years of MIR, most research focused on content-based image retrieval. Recently, there has been a surge of interest in a wide variety of media. An excellent example, life records, which encompasses all types of media simultaneously is being actively promoted by Bell [2004]. He is investigating the issues and challenges in processing life records—all the text, audio, video, and media related to a person's life.

Beyond text, audio, images, and video, there has been significant recent interest in new media such as 3D models. Assfalg et al. [2004] discuss using *spin-images*, which essentially encode the density of mesh vertices projected onto a 2D space, resulting in a 2D histogram. It was found that they give an effective view-independent representation for searching through a database of cultural artifacts. Funkhouser et al. [2003] develop a search engine for 3D models based on shape matching, using spherical harmonics to compute discriminating similarity measures which are effective even in the presence of model degeneracies. An overview of how 3D models are used in content-based retrieval systems can be found in Tangelder and Veltkamp [2004].

Another fascinating area is peering into biological databases consisting of imagery from the atomic through the visible light range. Applications can range from understanding the machinery of life to fast identification of dangerous bacteria or viruses. The aspect of particular interest is how to combine the data from different imaging methods such as electron microscopes, MRI, X-ray, and so on. Each imaging method uses a fundamentally different technique, however, the underlying content is the same. For example, Haas et al. [2004] used a genetic algorithm learning approach combined with additional knowledge sources to search through virus databases and video collections. To support imprecise queries in bio-databases, Chen et al. [2002] used fuzzy equivalence classes to assist query relaxation in biological imagery collections.

2.5 Browsing and Summarization

There have been a wide variety of innovative ways of browsing and summarizing multimedia information. Spierenburg and Huijsmans [1997] proposed a method for converting an image database into a movie. The intuition was that one could cluster a sufficiently large image database so that visually similar images would be in the same cluster. After the cluster process, one can order the clusters by the intercluster similarity, arrange the images in sequential order, and then convert to a video. This allows a user to have a gestalt understanding of a large image database in minutes.

Sundaram et al. [2002] took a similar approach toward summarizing video. They introduced the idea of a video skim which is a shortened video composed of informative scenes from the original video. The fundamental idea is for the user to be able to receive an abstract of the story but in video format.

Snoek et al. [2005] propose several methods for summarizing video such as grouping by categories and browsing by category and in time. Chiu et al. [2005] created a system for texturing a 3D city with relevant frames from video shots. The user would then be able to fly-through the 3D city and browse all of the videos in a directory. The most important frames would be located on the roofs of the buildings in the city so that a high altitude fly-through would result in viewing a single frame-per-video.

Uchihashi et al. [1999] suggested a method for converting a movie into a cartoon strip in the Manga style from Japan. This means altering the size and position of the relevant keyframes from the video based on their importance. Tian et al. [2002] took the concept of variable size and positions of images to the next level by posing the problem as a general optimization criterion problem, that is, what is the optimal arrangement of images on the screen so that the user can optimally browse an image database.

Liu et al. [2004] address the problem of effective summarization of images from WWW image search engines. They compare a rank list summarization method to an image clustering scheme and find that their users find the clustering scheme allows them to explore the image results more naturally and effectively.

2.6 High Performance Indexing

In the early multimedia database systems, the multimedia items such as images or video were frequently simply files in a directory or entries in an SQL database table. From a computational efficiency

perspective, both options exhibited poor performance because most filesystems use linear search within directories, and most databases could only perform efficient operations on fixed size elements. Thus, as the size of the multimedia databases or collections grew from hundreds, to thousands, to millions of variable sized items, the computers could not respond in an acceptable time period.

Even as the typical SQL database systems began to implement higher performance table searches, the search keys had to be exact such as in text search. Audio, images, and video were stored as blobs which could not be indexed effectively. Therefore, researchers [Egas et al. 1999; Lew 2000] turned to similarity-based databases which used tree-based indexes to achieve logarithmic performance. Even in the case of multimedia oriented databases such as the Informix database, it was still necessary to create custom datablades to handle efficient similarity searching such as k-d trees [Egas et al. 1999]. In general, the k-d tree methods had linear worst-case performance and logarithmic average-case performance in the context of feature-based similarity searches. A recent improvement to the k-d tree method is to integrate entropy-based balancing [Scott and Shyu 2003].

Other data representations have also been suggested besides k-d trees. Ye and Xu [2003] show that vector quantization can be used effectively for searching large databases. Elkwae and Kabuka [2000] propose a 2-tier signature-based method for indexing large image databases. Type 1 signatures represent the properties of the objects found in the images. Type 2 signatures capture the interobject spatial positioning. Together these signatures allow them to achieve a 98% performance improvement. Shao et al. [2003] use invariant features together with efficient indexing to achieve near real-time performance in the context of k nearest neighbor searching.

Other kinds of high performance indexing problems appear when searching peer to peer (P2P) networks due to the curse of dimensionality, the high communication overhead, and the fact that all searches within the network are based on nearest neighbor methods. Muller and Henrich [2003] suggest an effective P2P search algorithm based on compact peer data summaries. They show that their model allows peers to only communicate with a small sample and still retain high quality results.

2.7 Evaluation

Perhaps the most complete evaluation project in the last decade has been the TRECVID [Smeaton and Over 2003] evaluation. In TRECVID, there is a close connection between private industry and academic research where a realistic task-specific test set is gathered, discussed, agreed upon, and then numerous research teams attempt to provide the best video retrieval system for the test set. The main strengths are assembling a series of test collections for a certain type of user with a certain type of information need, and a set of relevance judgments on the topics for shots taken from the video reflecting a single real-world scenario. Test collections could include video with speech transcripts, machine translation of non-English speech, closed captions, metadata, common shot bounds, commonly used keyframes, and a set of automatically extracted features plus a set of multimedia topics (text, image, video). Important aspects have also been the process for creating realistic test sets, testing the research systems, and, most importantly, the continual evolution toward improving the test each year.

The most general recent work toward benchmarking has been on improving or completing performance graphs [Huijsmans and Sebe 2005]. They explain the limitations of the typical precision-recall graphs and develop additional performance indicators to cover the limitations. Normalization is suggested with respect to relevant class size and restriction to specific normalized scope values.

Keyframe-based retrieval techniques are the most popular in video retrieval systems. They represent a video as a small set of frames taken from the video content. Pickering and Ruger [2003] perform an evaluation of two learning methods (boosting and k-means) versus a vector space model. In the case of category searches, the k-means outperformed the other methods due to better handling of visually

disparate queries. The boosting algorithm performed best at finding video keyframes with similar compositions. Silva et al. [2005] discuss evaluation of video summarization for a large number of cameras in the context of a ubiquitous home. They implemented several keyframe extraction algorithms and found that using an adaptive algorithm-based on camera changes and footsteps, gives high quality results. Smeaton and Over [2003] discuss a complete evaluation of video retrieval systems which considers usage patterns and realistic test sets and compares a wide set of contributed systems.

Evaluation of multimedia retrieval has been an ongoing challenging problem [Huisman and Sebe 2005; Foote 1999; Downie 2003; Smeaton and Over 2003]. Audio, images, and video share distinct similarities such as the complex nature of content-based queries, overcoming intellectual property hurdles, and determination of what a reasonable person would find as relevant results. Furthermore, in the case of image retrieval, it has been shown that commonly used test databases such as the Corel stock image set are not necessarily effective performance indicators for real-world problems [Muller et al. 2002].

3. FUTURE DIRECTIONS

Despite the considerable progress of academic research in multimedia information retrieval, there has been relatively little impact of MIR research on commercial applications with some niche exceptions such as video segmentation. One example of an attempt to merge academic and commercial interests is Riya (www.riya.com). Riya's goal is to have a commercial product that uses the academic research in face detection and recognition and allows the users to search through their own photo collection or through the Internet for particular people. Another example is the MagicVideo Browser (www.magicbot.com) which transfers MIR research in video summarization to household desktop computers and has a plug-in architecture intended for easily adding new promising summarization methods as they appear in the research community. An interesting long-term initiative is the launching of Yahoo! Research Berkeley (research.yahoo.com/Berkeley), a new research partnership between Yahoo! Inc. and UC Berkeley whose declared scope is to explore and invent social media and mobile media technology and applications that will enable people to create, describe, find, share, and remix media on the Web. Nevenvision (www.nevenvision.com) is developing technology for mobile phones that utilizes visual recognition algorithms for bringing in ambient finding technology. However, these efforts are just in their infancy, and it is important to avoid a future where the MIR community is isolated from real-world interests. We believe that the MIR community has a golden opportunity in the growth of the multimedia search field that is commonly considered the next major frontier of search [Battelle 2005].

An issue in the collaboration between academic researchers and industry is the opaqueness of private industry. Frequently it is difficult to assess if commercial projects are using methods from the field of content-based MIR. In the current atmosphere of intellectual property lawsuits, many companies are reluctant to publish the details of their systems in open academic circles for fear of being served with a lawsuit. Nondisclosure can be a protective shield, but it does impede open scientific progress. This is a small hurdle if the techniques developed by researchers have significant direct application to practical systems.

To assess research effectively in multimedia retrieval, task-related standardized databases on which different groups can apply their algorithms are needed. In text retrieval, it has been relatively straightforward to obtain large collections of old newspaper texts because the copyright owners do not see the raw text as having much value. However image, video, and speech libraries do see great value in their collections and consequently are much more cautious in releasing their content. While it is not a research challenge, obtaining large multimedia collections for widespread evaluation benchmarking is a practical and important step that needs to be addressed. One possible solution is to see that task-related image and video databases with appropriate relevance judgments are included and made available to

groups for research purposes as was done with TRECVID. Useful video collections could include news video (in multiple languages), collections of personal videos, and possibly movie collections. Image collections would include image databases (maybe on specific topics) along with annotated text—(the use of library image collections should also be explored). One critical point here is that sometimes the artificial collections like Corel might do more harm than good to the field by misleading people into believing that their techniques work, while they do not necessarily work with more general image collections.

Therefore, cooperation between private industry and academia is strongly encouraged. The key point here is to focus on efforts which mutually benefit both industry and academia. As was noted earlier, it is of clear importance to keep in mind the needs of the users in retrieval system design, and it is logical that industry can contribute substantially to our understanding of the end-user and also aid in the realistic evaluation of research algorithms. Furthermore, by having closer communication with private industry, we can potentially find out what parts of their systems need additional improvements to increase user satisfaction. In the example of Riya, they clearly need to perform object detection (faces) on complex backgrounds and then object recognition (who the face is). In the context of consumer digital photograph collections, the MIR community might attempt to create a solid test set which could be used to assess the efficacy of different algorithms in both detection and recognition in real-world media.

The potential landscape of multimedia information retrieval is quite wide and diverse. Following are some potential areas for additional MIR research challenges.

Human Centered Methods. We should focus as much as possible on the user who may want to explore instead of search for media. It has been noted that decision makers need to explore an area to acquire valuable insight, thus experiential systems which stress the exploration aspect are strongly encouraged. Studies on the needs of the user are also highly encouraged to give us a full understanding of their patterns and desires. New interactive devices (e.g., force, olfactory, and facial expression detectors) have largely been overlooked and should be tested to provide new possibilities such as human emotional state detection and tracking.

Multimedia Collaboration. Discovering more effective means of human-human computer-mediated interaction is increasingly important as our world becomes more wired or wirelessly connected. In a multimodal collaboration environment, many questions remain: How do people find one another? How does an individual discover meetings/collaborations? What are the most effective multimedia interfaces in these environments for different purposes, individuals, and groups? Multimodal processing has many potential roles ranging from transcribing and summarizing meetings to correlating voices, names, and faces, to tracking individual (or group) attention and intention across media. Careful and clever instrumentation and evaluation of collaboration environments will be key to learning more about just how people collaborate.

Very important here is the query model which should benefit from the collaboration environment. One solution would be to use an event-based query approach [Liu et al. 2005] that can provide the users a more feasible way to access the related media content with the domain knowledge provided by the environment model. This approach could be extremely important when dealing with live multimedia where the multimedia information is captured in a real-life setting by different sensors and streamed to a central processor.

Interactive Search and Agent Interfaces. Emergent semantics and its special case of relevance feedback methods are quite popular because they potentially allow the system to learn the goals of the user in an interactive way. Another perspective is that relevance feedback is serving as a special type of smart *agent interface*. Agents are present in learning environments, games, and customer service applications. They can mitigate complex tasks, bring expertise to the user, and provide more natural interaction. For example, they might be able to adapt sessions to a user, deal with dialog interruptions or follow-up questions, and help manage the focus of attention. Agents raise important technical and

social questions but equally provide opportunities for research in representing, reasoning about, and realizing agent belief and attitudes (including emotions). Creating natural behaviors and supporting speaking and gesturing agent displays are important user interface requirements. Research issues include what the agents can and should do, how and when they should do it (e.g., implicit versus explicit tasking, activity, and reporting), and by what means they should carry out communications (e.g., text, audio, video). Other important issues include how do we instruct agents to change their future behavior, and who is responsible when things go wrong.

Neuroscience and New Learning Models. Observations of child learning and neuroscience suggest that exploiting information from multiple modalities (i.e., audio, imagery, haptic) reduces processing complexity. For example, researchers have begun to explore early word acquisition from natural acoustic descriptions and visual images (e.g., shape, color) of everyday objects in which mutual information appears to dramatically reduce computational complexity. This work, which exploits results from speech processing, computer vision, and machine learning, is being validated by observing mothers as play with their prelinguistic infants performing the same task. Neuroscientists and cognitive psychologists are only beginning to discover and, in some cases, validate abstract functional architectures of the human mind. However, even the relatively abstract models available from today's measurement techniques (e.g., low fidelity measures of gross neuroanatomy via indirect measurement of neural activity such as cortical blood flow) promise to provide us with new insight and inspire innovative processing architectures and machine learning strategies.

Caution should be used when such neuroscience-inspired models are considered. These models are good for inspiration and high-level ideas. However, they should not be carried too far because the computational machinery is very different. The neuroscience/cognition community tries to form the model of a human machine, and we are trying to develop tools that will be useful for humans. There is some overlap, but the goals are rather different.

In general, there is great potential in tapping into or collaborating with the artificial intelligence and learning research community for new paradigms and models of which neuro-based learning is only one candidate. Learning methods have great potential for synergistically combining multiple media at different levels of abstraction. Note that the current search engines (e.g., Yahoo!, Google, etc) use only text for indexing images and video. Therefore, approaches which demonstrate synergy of text with image and video features have significant potential. Note that learning must be applied at the right level as is done in some hierarchical approaches and also in the human brain. An arbitrary application of learning might result in techniques that are very fragile and are useless except for some niche cases. Furthermore, services such as Blinkx and Riya currently utilize learning approaches to extract words in movies from complex, noisy audio tracks (Blinkx) or detecting and recognizing faces from photos with complex backgrounds (Riya). In both cases, only methods which are robust to the presence of real-world noise and complexity will be beneficial in improving the effectiveness of similar services.

Folksonomies. It is clear that the problem of automatically extracting content multimedia data is a difficult problem. Even in text, we could not do it completely. As a consequence, all the existing search engines are using simple keyword-based approaches or are developing approaches that have a significant manual component and address only specific areas. Another interesting finding is that, for an amorphous and large collection of information, a taxonomy-based approach could be too rigid for navigation. Since it is relatively easier to develop inverted file structures to search for keywords in large collections, people find the idea of tags attractive: by somehow assigning tags, we can organize relatively unstructured files and search. About the same time, the idea of the wisdom of crowd became popular. So it is easy to argue that tags could be assigned by people and will result in *wise tags* (because they are assigned by the crowd) and this will be a better approach than the dictatorial taxonomy. The idea is appealing and made flickr.com and Del.icio.us useful and popular.

The main question arises: Is this approach really working—or can it be made to work? If everybody assigns several appropriate tags to a photo and then the crowd seeing that photo also assigns appropriate tags, then the wisdom of crowd may come into action. But if the uploader rarely assigns tags, and the viewers, if any, assign tags even more rarely, then there is no crowd, and there is no wisdom. Interesting game-like approaches (see, e.g., www.espgame.org) are being developed to assign these tags to images. Based on ad hoc analysis, it seems that very few tags are being assigned to photos on flickr.com by people who upload images and fewer are being assigned by the viewers. Moreover, it may happen that, without any guidance, people become confused about how to assign tags. It appears that the success may come from some interesting combination of taxonomy and folksonomy.

No Solved Problems. From the most recent panel discussions at the major MIR scientific conferences including ACM MIR and CIVR, it is generally agreed that there are no solved problems. In some cases, a general problem is reduced to a smaller niche problem where high accuracy and precision can be quantitatively demonstrated, but the general problem remains largely unsolved. In summary, all of the general problems need significant further research.

4. MAJOR CHALLENGES

In conclusion, these major research challenges are noteworthy and of particular importance to the MIR research community: (1) semantic search with emphasis on the detection of concepts in media with complex backgrounds; (2) multimodal analysis and retrieval algorithms especially to exploit the synergy between the various media, including text and context information; (3) experiential multimedia exploration systems to allow users to gain insight and explore media collections; (4) interactive search, emergent semantics, or relevance feedback systems; and (5) evaluation with emphasis on representative test sets and usage patterns.

ACKNOWLEDGMENTS

We would like to thank Alberto del Bimbo, Shih-Fu Chang, Nevenka Dimitrova, Theo Gevers, William Grosky, Thomas Huang, John Kender, Lawrence Rowe, Alan Smeaton, and Arnold Smeulders for excellent discussions on the future of MIR.

REFERENCES

Amir, A., Basu, S., Iyengar, G., Lin, C.-Y., Naphade, M., Smith, J. R., Srinivasan, S., and Tseng, B. 2004. A Multi-modal system for the retrieval of semantic video events. *Comput. Vision Image Understand. 96,* 2, 216–236.

Assfalg, J., Del Bimbo, A., and Pala, P. 2004. Retrieval of 3D objects by visual similarity. In *Proceedings of the 6th International Workshop on Multimedia Information Retrieval*. New York, NY. (Oct.). M. S. Lew, N. Sebe, C. Djeraba, Eds. ACM, New York, NY. 77–83.

Bach, J. R., Fuller, C., Gupta, A., Hampapur, A., Horowitz, B., Humphrey, R., Jain, R., and Shu, C. F. 1996. Virage image search engine: An open framework for image management. In *Proceedings of the SPIE Storage and Retrieval for Still Image and Video Databases*. 76–87.

Balakrishnan, N., Hariharakrishnan, K., and Schonfeld, D. 2005. A new image representation algorithm inspired by image submodality models, redundancy reduction, and learning in biological vision. *IEEE Trans. Patt. Analy. Machine Intellig. 27*, 9, 1367–1378.

Ballard, D. H. and Brown, C. M. 1982. *Computer Vision*. Prentice Hall, New Jersey, USA.

Bakker, E. M. and Lew, M. S. 2002. Semantic video retrieval using audio analysis. In *Proceedings of the 1st International Conference on Image and Video Retrieval*. (July) London, UK. M. S. Lew, N. Sebe, and J. P. Eakins, Eds. Springer-Verlag, London, UK. 262–270.

Bartolini, I., Ciaccia, P., and Patella, M. 2005. WARP: Accurate retrieval of shapes using phase of fourier descriptors and time warping distance. *IEEE Trans. Patt. Analy. Machine Intellig. 27*, 1, 142–147.

Battelle, J. 2005. *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*. Portfolio Hardcover.

Baumberg, A. 2000. Reliable feature matching across widely separated views. *IEEE Conference of Computer Vision and Pattern Recognition*. 774–781.

Bell, G. 2004. A new relevance for multimedia when we record everything personal. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*. ACM, New York, NY.

Benitez, A. B. and Chang, S.-F. 2002. Semantic knowledge construction from annotated image collection. In *Proceedings of the IEEE International Conference on Multimedia*. IEEE Computer Society Press, Los Alamitos, CA.

Beretti, S., Del Bimbo, A., and Vicario, E. 2001. Efficient matching and indexing of graph models in content-based retrieval. *IEEE Trans. Patt. Analy. Machine Intellig. 23*, 10, 1089–1105.

Berthouze, N. B. and Kato, T. 1998. Towards a comprehensive integration of subjective parameters in database browsing. In *Advanced Database Systems for Integration of Media and User Environments*, Y. Kambayashi, A. Makinouchi, S. Uemura, K. Tanaka, and Y. Masunaga, Eds. World Scientific, Singapore, 227–232.

Bliujute, R., Saltenis, S., Slivinskas, G., and Jensen, C. S. 1999. Developing a DataBlade for a new index. In *Proceedings of IEEE International Conference on Data Engineering*. (March) Sydney, Australia 314–323.

Bosson, A., Cawley, G. C., Chan, Y., and Harvey, R. 2002. Non-retrieval: Blocking Pornographic Images. In *Proceedings of the 1st International Conference on Image and Video Retrieval*. (July) London, M. S. Lew, N. Sebe, and J. P. Eakins, Eds. Springer-Verlag, London, UK, 50–60.

Cappelli, R., Maio, D., and Maltoni, D. 2001. Multispace KL for pattern representation and classification. *IEEE Trans. Pattern Analy. Machine Intellig. 23*, 9, 977–996.

Chang, S.-F., Chen, W., and Sundaram, H. 1998. Semantic visual templates: Linking visual features to semantics. In *Proceedings of the IEEE International Conference on Image Processing*. IEEE Computer Society Press, Los Alamitos, CA. 531–535.

Chen, Y., Che, D., and Aberer, K. 2002. On the efficient evaluation of relaxed queries in biological databases. In *Proceedings of the 11th International Conference on Information and Knowledge Management*. McLean, VA, 227–236.

Chen, Y., Zhou, X. S., and Huang, T. S. 2001. One-class SVM for learning in image retrieval. In *Proceedings of IEEE International Conference on Image Processing*, (Oct.), Thessaloniki, Greece, 815–818.

Chiu, P., Girgensoh, A., Lertsithichai, S., Polak, W., and Shipman, F. 2005. MediaMetro: Browsing multimedia document collections with a 3D city metaphor. In *Proceedings of the 13th ACM International Conference on Multimedia*. (Nov.), Singapore, 213–214.

Chua, T. S., Zhao, Y., and Kankanhalli, M. S. 2002. Detection of human faces in a compressed domain for video stratification. *The Visual Computer 18*, 2, 121–133.

Cooper, M., Foote, J., Girgensohn, A., and Wilcox, L. 2005. Temporal event clustering for digital photo collections. *ACM Trans. Multimedia Comput. Comm. Applica. 1*, 3, 269–288.

Dimitrova, N., Agnihotri, L., and Wei, G. 2000. Video classification based on HMM using text and faces. *European Signal Processing Conference*. Tampere, Finland.

Dimitrova, N., Zhang, H. J., Shahraray, B., Sezan, I., Huang, T., and Zakhor, A. 2002. Applications of video-content analysis and retrieval. *IEEE Multimedia 9*, 3, 42–55.

Dimitrova, N. 2003. Multimedia content analysis: The next wave. In *Proceedings of the 2nd International Conference on Image and Video Retrieval*. (July), Urbana, E. M. Bakker, T. S. Huang, M. S. Lew, N. Sebe, and X. Zhou, Eds. Springer-Verlag, London, UK, 9–18.

Djeraba, C. 2002. Content-based multimedia indexing and retrieval. *IEEE Multimedia 9*, 18–22.

Djeraba, C. 2003. Association and content-based retrieval. *IEEE Trans. Knowl. Data Engin. 15*, 1, 118–135.

Downie, J. S. 2003. Toward the scientific evaluation of music information retrieval systems. In *Proceedings of the International Conference on Music Information Retrieval*. Baltimore, MD, 25–32.

Dufournaud, Y., Schmid, C., and Horaud, R. 2000. Matching images with different resolutions. *IEEE Conference of Computer Vision and Pattern Recognition*. 612–618.

Dy, J. G., Brodley, C. E., Kak, A., Broderick, L. S., and Aisen, A. M. 2003. Unsupervised feature selection applied to content-based retrieval of lung images. *IEEE Trans. Patt. Analy. Machine Intellig. 25*, 3, 373–378.

Eakins, J. P., Riley, K. J., and Edwards, J. D. 2003. Shape feature matching for trademark image retrieval. In *Proceedings of the 2nd International Conference on Image and Video Retrieval*. (July), Urbana, IL. E. M. Bakker, T. S. Huang, M. S. Lew, N. Sebe, and X. Zhou, Eds. Springer-Verlag, London, UK. 28–38.

Egas, R., Huijsmans, N., Lew, M. S., and Sebe, N. 1999. Adapting k-d Trees to Visual Retrieval. In *Proceedings of the International Conference on Visual Information Systems*. (June) Amsterdam, A. Smeulders and R. Jain, Eds., 533–540.

Eiter, T. and Libkin, L. 2005. *Database Theory*. Springer, London. UK.

Elkwae, E. A. and Kabuka, M. R. 2000. Efficient content-based indexing of large image databases. *ACM Trans. Inform. Sys. 18*, 2, 171–210.

ENSER, P. G. B. AND SANDOM, C. J. 2003. Towards a comprehensive survey of the semantic gap in visual information retrieval. In *Proceedings of the 2nd International Conference on Image and Video Retrieval*. (July), Urbana, IL, E. M. Bakker, T. S. Huang, M. S. Lew, N. Sebe and X. Zhou, Eds. Springer-Verlag, London, UK. 291–299.

ENSER, P. G. B., SANDOM, C. J., AND LEWIS, P. H. 2005. Automatic annotation of images from the practitioner perspective. In *Proceedings of the 4th International Conference on Image and Video Retrieval*. (July) Singapore, IL W. Leow, M. S. Lew, T.-S. Chua, W.-Y. Ma, E. M. Bakker, and L. Chaisorn, Eds. Springer-Verlag, London, UK. 497–506.

FAN, J., GAO, Y., AND LUO, H. 2004. Multi-level annotation of natural scenes using dominant image components and semantic concepts. In *Proceedings of the ACM International Conference on Multimedia*. ACM, New York, NY, 540–547.

FLICKNER, M., SAWHNEY, H., NIBLACK, W., ASHLEY, J., QIAN HUANG DOM, B., GORKANI, M., HAFNER, J., LEE, D., PETKOVIC, D., STEELE, D., AND YANKER, P. 1995. Query by image and video content: The QBIC system. *IEEE Comput.*, (Sept.), 23–32.

FOOTE, J. 1999. An overview of audio information retrieval. *ACM Multimedia Syst. 7*, 1, 42–51.

FOOTE, J. 2000. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. IEEE Computer Society Press, Los Alamitos, CA, 452–455.

FORSYTH, D. A. AND FLECK, M. M. 1999. Automatic detection of human nudes. *Int. J. Comput. Vision 32*, 1, 63–77.

FRANKEL, C., SWAIN, M. J., AND ATHITSOS, V. 1996. WebSeer: An image search engine for the World Wide Web. University of Chicago Tech. rep. 96-14, University of Chicago, Chicago, IL.

FROHLICH, D., KUCHINSKY, A., PERING, C., DON, A., AND ARISS, S. 2002. Requirements for photoware. In *Proceedings of the ACM Conference on CSCW*. ACM Press, New York, NY, 166–175.

FUNKHOUSER, T., MIN, P., KAZHDAN, M., CHEN, J., HALDERMAN, A., DOBKIN, D., AND JACOBS, D. 2003. A search engine for 3D models. *ACM Trans. Graph. 22*, 1, 83–105.

GEVERS, T. 2001. Color-based retrieval. In *Principles of Visual Information Retrieval*, M. S. Lew, Ed. Springer-Verlag, London, UK, 11–49.

GONG, B., SINGH, R., AND JAIN, R. 2004. ResearchExplorer: Gaining insights through exploration in multimedia scientific data. In *Proceedings of the 6th International Workshop on Multimedia Information Retrieval*. (Oct.) New York, M. S. Lew, N. Sebe, C. Djeraba, Eds. ACM, New York, NY, 7–14.

GRAHAM, A., GARCIA-MOLINA, H., PAEPCKE, A., AND WINOGRAD, T. 2002. Time as the essence for photo browsing through personal digital libraries. In *Proceedings of the Joint Conference on Digital Libraries*. ACM Press, New York, NY, 326–335.

GREENSPAN, H., GOLDBERGER, J., AND MAYER, A. 2004. Probabilistic space-time video modeling via piecewise GMM. *IEEE Trans. Patt. Analy. Machine Intell. 26*, 3, 384–396.

GUO, G., ZHANG, H. J., AND LI, S. Z. 2001. Boosting for content-based audio classification and retrieval: An Evaluation, In *Proceedings of the IEEE Conference on Multimedia and Expo*. (Aug.) Tokyo, Japan.

HAAS, M., LEW, M. S., AND HUIJSMANS, D. P. 1997. A new method for key frame based video content representation. In *Image Databases and Multimedia Search*. A. Smeulders and R. Jain, Eds. World Scientific. 191–200.

HAAS, M., RIJSDAM, J., AND LEW, M. 2004. Relevance feedback: Perceptual learning and retrieval in bio-computing, photos, and video, In *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*. (Oct.), New York, 151–156.

HANJALIC, A., LAGENDIJK, R. L., AND BIEMOND, J. 1997. A new method for key frame based video content representation. In *Image Databases and Multimedia Search*, A. Smeulders and R. Jain, Eds. World Scientific. 97–107.

HANJALIC, A. AND XU, L.-Q. 2005. Affective video content representation and modeling. *IEEE Trans. Multimedia, 7*, 1, 171–180.

HARALICK, R. M. AND SHAPIRO, L. G. 1993. *Computer and Robot Vision*. Addison-Wesley, New York, NY.

HARRIS, C. AND STEPHENS, M. 1988, A combined corner and edge detector. *The 4th Alvey Vision Conference*. 147–151.

HASTINGS, S. K. 1999. Evaluation of image retrieval Systems: Role of User Feedback. *Library Trends 48*, 2, 438–452.

HE, X., MA, W.-Y., KING, O., LI, M., AND ZHANG, H. 2002. Learning and inferring a semantic space from user's relevance feedback for image retrieval. In *Proceedings of the ACM Multimedia Conference*. ACM, New York, NY, 343–347.

HOWE, N. 2003. A closer look at boosted image retrieval. *In Proceedings of the 2nd International Conference on Image and Video Retrieval*. (July), Urbana, IL, E. M. Bakker, T. S. Huang, M. S. Lew, N. Sebe, and X. Zhou, Eds. Springer-Verlag, London, UK, 61–70.

HUIJSMANS, D. P. AND SEBE, N. 2005. How to complete performance graphs in content-based image retrieval: Add generality and normalize Scope. *IEEE Trans. Patt. Analy. Machine Intellig. 27*, 2, 245–251.

JACOBS, D. W., WEINSHALL, D., AND GDALYAHU, Y. 2000. Classification with nonmetric distances: Image hetrieval and class representation. *IEEE Trans. Patt. Analy. Machine Intell. 22*, 6, 583–600.

JAFARI-KHOUZANI, K. AND SOLTANIAN-ZADEH, H. 2005. Radon transform orientation estimation for rotation invariant texture analysis. *IEEE Trans. Patt. Analy. Machine Intell. 27*, 6, 1004–1008.

JAIMES, A. AND SEBE, N. 2006. Multimodal human-computer interaction: A survey. *Comput. Vision Image Understand.* To appear.

JAIN, R. 2003. A game experience in every application: Experiential computing. *Comm. ACM 46*, 7, 48–54.

JAIN, R., KIM, P., AND LI, Z. 2003. Experiential meeting system. In *Proceedings of the 2003 ACM SIGMM Workshop on Experiential Telepresence*. Berkeley, CA, 1–12.

JOLION, J. M. 2001. Feature similarity. In *Principles of Visual Information Retrieval*. M. S. Lew, Ed. Springer-Verlag, London, UK. 122–162.

KRISHNAPURAM, R., MEDASANI, S., JUNG, S. H., CHOI, Y. S., AND BALASUBRAMANIAM, R. 2004. Content-based image retrieval based on a fuzzy approach. *IEEE Trans. Knowl. Data Eng. 16*, 10, 1185–1199.

LEVINE, M. 1985. *Vision in Man and Machine*. Mcgraw Hill, Columbus, OH.

LEW, M. S. AND HUIJSMANS, N. 1996. Information theory and face detection. In *Proceedings of the International Conference on Pattern Recogntion*. Vienna, Austria, 601–605.

LEW, M. S. 2000. Next generation Web searches for visual content. *IEEE Comput.* (Nov.). 46–53.

LEW, M. S. 2001. *Principles of Visual Information Retrieval*. Springer, London, UK.

LEW, M. S. AND DENTENEER, D. 2001. Fisher keys for content based retrieval. *Image Vision Comput. 19*, 561–566.

LI, J. AND WANG, J. Z. 2003. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Patt. Analy. Machine Intell. 25*, 9, 1075–1088.

LIENHART, R. 2001. Reliable transition detection in videos: A survey and practitioner's guide. *Int. J. Image Graph. 1*, 3, 469–486.

LIM, J.-H., TIAN, Q., AND MULHELM, P. 2003. Home photo content modeling for personalized event-based retrieval. *IEEE Multimedia 10*, 4, 28–37.

LINDEBERG, T. 1998. Feature detection with automatic scale selection. *Int. J. Comput. Vision*. *30*, 2, 79–116.

LINDEBERG, T. AND GARDING, J. 1997. Shape-adapted smoothing in estimation of the 3D shape cues from affine deformations of local 2D brightness structure. *Image Vision Comput. 15*, 6, 415–434.

LIU, B., GUPTA, A., AND JAIN, R. 2005. MedSMan: A streaming data management system over live multimedia. *ACM Multimedia*, 171–180.

LIU, H., XIE, X., TANG, X., LI, Z. W., AND MA, W. Y. 2004. Effective browsing of Web image search results. In *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*. ACM, New York, NY. 84–90.

LIU, X., SRIVASTAVA, A., AND SUN, D. 2003. Learning optimal representations for image retrieval applications. In *Proceedings of the 3rd International Conference on Image and Video Retrieval*. (July), Urbana, IL, E. M. Bakker, T. S. Huang, M. S. Lew, N. Sebe, and X. Zhou, Eds. Springer-Verlag, London, UK. 50–60.

LOWE, D. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision 60*, 2, 91–110.

MARKKULA, M. AND SORMUNEN, E. 2000. End-user searching challenges indexing practices in the digital newspaper photo archive. *Inform. Retrieval 1*, 4, 259–285.

MIKOLAJCZYK, K. AND SCHMID, C. 2004, Scale and affine invariant interest point detectors. *Int. J. Comput. Vision 60*, 1, 63–86.

MONGY, S., BOUALI, F., AND DJERABA, C. 2005. Analyzing user's behavior on a video database. In *Proceedings of ACM MDM/KDD Workshop on Multimedia Data Mining*. Chicago, IL.

MULLER, H., MULLER, W., MARCHAND-MAILLET, S., PUN, T., AND SQUIRE, D. 2000. Strategies for positive and negative relevance feedback in image retrieval. In *Proceedings of 15th International Conference on Pattern Recognition*. (Sept.) Barcelona, Spain, 1043–1046.

MULLER, H., MARCHAND-MAILLET, S., AND PUN, T. 2002. The Truth about Corel-evaluation in image retrieval. In *Proceedings of the 1st International Conference on Image and Video Retrieval*. (July), London, UK, M. S. Lew, N. Sebe, and J. P. Eakins, Eds. Springer-Verlag, London, UK. 38–49.

MÜLLER, W. AND HENRICH, A. 2003. Fast retrieval of high-dimensional feature vectors in P2P networks using compact peer data summaries. In *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval*, Berkeley, CA, 79–86.

OJALA, T., PIETIKAINEN, M., AND HARWOOD, D. 1996. Comparative study of texture measures with classification based on feature distributions. *Patt. Recogn. 29*, 1, 51–59.

PEREIRA, F. AND KOENEN, R. 2001. MPEG-7: A standard for multimedia content description. *Int. J. Image Graph. 1*, 3, 527–546.

PICARD, R. W. 2000. *Affective Computing*. MIT Press, Cambridge, MA.

PICKERING, M. J. AND RÜGER, S. 2003. Evaluation of key-frame based retrieval techniques for video. *Comput. Vision Image Understand. 92*, 2, 217–235.

RAUTIAINEN, M., SEPPANEN, T., PENTTILA, J., AND PELTOLA, J. 2003. Detecting semantic concepts from video using temporal gradients and audio classification. In *Proceedings of the 3rd International Conference on Image and Video Retrieval*. (July), Urbana, IL, E. M. Bakker, T. S. Huang, M. S. Lew, N. Sebe, and X. Zhou, Eds. Springer-Verlag, London, UK, 260–270.

ROCCHIO 1971. Relevance feedback in information retrieval. In *The Smart Retrieval System: Experiments in Automatic Document Processing*. G. Salton, Ed. Prentice Hall, Englewoods Cliffs, NJ.

RODDEN, K., BASALAJ, W., SINCLAIR, D., AND WOOD, K. 2001. Does organisation by similarity assist image browsing? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. (Mar.), Seattle, WA. 190–197.

RODDEN, K. AND WOOD, K. 2003. How do people manage their digital photographs? In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM Press, New York, NY, 409–416.

ROWE, L. A. AND JAIN, R. 2005. ACM SIGMM retreat report on future directions in multimedia research. *ACM Trans. Multimedia Comput. Comm. Appl. 1*, 1, 3–13.

ROWLEY, H., BALUJA, S., AND KANADE, K. 1996. Human face detection in visual scenes. In *Proceedings of NIPS Advances in Neural Information Processing Systems 8*, (Nov.), Denver, CO, 875–881.

RUBIN, R. 2004. *Foundations of Library and Information Science*. Neal-Schuman Publishers, New York, NY.

RUI, Y. AND HUANG, T. S. 2001. Relevance feedback techniques in image retrieval. In *Principles of Visual Information Retrieval*, M. S. Lew, Ed. Springer-Verlag, London, UK, 219–258.

SALWAY, A. AND GRAHAM, M. 2003. Extracting information about emotions in films. In *Proceedings of the ACM International Conference on Multimedia*. (Nov.) Berkeley, CA, 299–302.

SCHMID, C., MOHR, R., AND BAUCKAGE, C. 2000, Evaluation of interest point detectors. *Int. J. Comput. Vision 37*, 2, 151–172.

SCHNEIDERMAN, H. AND KANADE, T. 2004. Object detection using the statistics of parts. *Int. J. Comput. Vision 56*, 3, 151–177.

SCLAROFF, S., LA CASCIA, M., SETHI, S., AND TAYCHER, L. 2001. Mix and match features in the ImageRover search engine. In *Principles of Visual Information Retrieval*. M. S. Lew, Ed. Springer-Verlag, London, UK, 259–277.

SCOTT, G. J. AND SHYU, C. R. 2003. EBS k-d tree: An entropy balanced statistical k-d tree for image databases with ground-truth labels. In *Proceedings of the 2nd International Conference on Image and Video Retrieval*. (July), Urbana, IL, E. M. Bakker, T. S. Huang, M. S. Lew, N. Sebe, and X. Zhou, Eds. Springer-Verlag, London, UK, 467–476.

SEBASTIAN, T. B., KLEIN, P. N., AND KIMIA, B. B. 2004. Recognition of shapes by editing their shock graphs. *IEEE Trans. Patt. Analy. Machine Intell. 26*, 5, 550–571.

SEBE, N., LEW, M. S., AND HUIJSMANS, D. P. 2000. Toward improved ranking metrics. *IEEE Trans. Patt. Analy. Mach. Intell. 22*, 10, 1132–1143.

SEBE, N. AND LEW, M. S. 2001. Color based retrieval. *Pattern Recognition Letters 22*, 2, 223–230.

SEBE, N. AND LEW, M. S. 2002. Robust shape matching. In *Proceedings of the 1st International Conference on Image and Video Retrieval*. (July) London, UK, M. S. Lew, N. Sebe, and J. P. Eakins, Eds. Springer-Verlag, London, UK, 17–28.

SEBE, N., COHEN, I., GARG, A., LEW, M. S., AND HUANG, T. S. 2002. Emotion recognition using a Cauchy naive Bayes classifier. In *Proceedings of International Conference on Pattern Recognition*. (Aug.) Quebec, Canada, 17–20.

SEBE, N., TIAN, Q., LOUPIAS, E., LEW, M. S., AND HUANG, T. S. 2003a. Evaluation of salient point techniques. *Image Vision Computing 21*, 13–14, 1087–1095.

SEBE, N., LEW, M. S., ZHOU, X., AND HUANG, T. S. 2003b. The state of the art in image and video retrieval. In *Proceedings of the 2nd International Conference on Image and Video Retrieval*. (July), Urbana, IL, E. M. Bakker, T. S. Huang, M. S. Lew, N. Sebe, and X. Zhou, Eds. Springer-Verlag, London, UK.

SHAO, H., SVOBODA, T., TUYTELAARS, T., AND VAN GOOL, L. 2003. HPAT indexing for fast object/scene recognition based on local appearance. In *Proceedings of the 2nd International Conference on Image and Video Retrieval*. (July), Urbana, IL, E. M. Bakker, T. S. Huang, M. S. Lew, N. Sebe, and X. Zhou, Eds. Springer-Verlag, London, UK, 71–80.

SHEN, H. T., OOI, B. C., AND TAN, K. L. 2000. Giving meanings to WWW images. In *Proceedings of ACM Multimedia*. ACM, New York, NY, 39–48.

SILVA, G. C., DE, YAMASAKI, T., AND AIZAWA, K. 2005. Evaluation of video summarization for a large number of cameras in ubiquitous home. In *Proceedings of the 13th ACM International Conference on Multimedia*. (Nov.) ACM, Singapore, 820–828.

SMEATON, A. F. AND OVER, P. 2003. Benchmarking the effectiveness of information retrieval tasks on digital video. In *Proceedings of the 2nd International Conference on Image and Video Retrieval*. (July), Urbana, IL, E. M. Bakker, T. S. Huang, M. S. Lew, N. Sebe, and X. Zhou, Eds. Springer-Verlag, London, UK, 10–27.

SMEULDERS, A., WORRING, M., SANTINI, S., GUPTA, A., AND JAIN, R. 2000. Content based image retrieval at the end of the early years. *IEEE Trans. Patt. Analy. Mach. Intell. 22*, 12, 1349–1380.

SMITH, J. R. AND CHANG, S. F. 1997. Visually searching the web for content. *IEEE Multimedia 4*, 3, 12–20.

Snoek, C. G. M., Worring, M., Van Gemert, J., Geusebroek, J. M., Koelma, D., Nguyen, G. P., De Rooij, O., and Seinstra, F. 2005. MediaMill: Exploring news video archives based on learned semantics. In *Proceedings of the 13th ACM International Conference on Multimedia*. (Nov.) Singapore, 225–226.

Spierenburg, J. A. and Huijsmans, D. P. 1997. VOICI: Video overview for image cluster indexing. In *Proceedings of the 8th British Machine Vision Conference*. (June) Colchester, UK.

Srivastava, A., Joshi, S. H., Mio, W., and Liu, X. 2005. Statistical shape analysis: Clustering, learning, and testing. *IEEE Trans. Patt. Analy. Mach. Intell. 27*, 4, 590–602.

Sundaram, H., Xie, L., and Chang, S. F. 2002. A utility framework for the automatic generation of audio-visual skims. In *Proceedings of the 10th ACM International Conference on Multimedia*. Juan-les-Pins, France, 189–198.

Tangelder, J. and Veltkamp, R. C. 2004. A survey of content based 3d shape retrieval methods. In *Proceedings of the International Conference on Shape Modeling and Applications*. (June) Genova, Italy. IEEE, New York, NY, 157–166.

Tian, Q., Sebe, N., Lew, M. S., Loupias, E., and Huang, T. S. 2001. Image retrieval using wavelet-based salient points. *Journal of Electronic Imaging 10*, 4, 835–849.

Tian, Q., Moghaddam, B., and Huang, T. S. 2002. Visualization, estimation and user-modeling. In *Proceedings of the 1st International Conference on Image and Video Retrieval*. (July), London, UK, M. S. Lew, N. Sebe, and J. P. Eakins, Eds. Springer-Verlag, London, UK, 7–16.

Tieu, K. and Viola, P. 2004. Boosting image retrieval. *Int. J. Comput. Vision 56*, 1, 17–36.

Therrien, C. W. 1989. *Decision, Estimation, and Classification*. Wiley, New York, NY.

Tuytelaars, T. and van Gool, L. 2000. Wide baseline stereo matching based on local affinely invariant regions. *British Machine Vision Conference*. 412–425.

Uchihashi, S., Foote, J., Girgensohn, A., and Boreczky, J. 1999. Video manga: Generating semantically meaningful video summaries. In *Proceedings of the 7th ACM International Conference on Multimedia*. Orlando, FL, 383–392.

Vailaya, A., Jain, A., and Zhang, H. 1998. On image classification: City vs landscape. In *Proceedings of Workshop on Content-Based Access of Image and Video Libraries*. 3–8.

Veltkamp, R. C. and Hagedoorn, M. 2001. State of the art in shape matching. In *Principles of Visual Information Retrieval*. M. S. Lew, Ed. Springer-Verlag, London, UK, 87–119.

Wang, W., Yu, Y., and Zhang, J. 2004. Image emotional classification: Static vs. dynamic. In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*. (Oct.), 6407–6411.

Winston, P. 1992. *Artificial Intelligence*, Addison-Wesley, New York, NY.

Worring, M. and Gevers, T. 2001. Interactive retrieval of color images. *Int. J. Image Graph. 1*, 3, 387–414.

Worring, M., Nguyen, G. P., Hollink, L., Gemert, J. C., and Koelma, D. C. 2004. Accessing video archives using interactive search. In *Proceedings of IEEE International Conference on Multimedia and Expo.* (June) IEEE, Taiwan, 297–300.

Wu, P., Choi, Y., Ro., Y. M., and Won, C. S. 2001. MPEG-7 texture descriptors. *Int. J. Image Graph. 1*, 3, 547–563.

Yang, M. H., Kriegman, D. J., and Ahuja. N. 2002. Detecting faces in images: A survey. *IEEE Trans. Patt. Analy. Machine Intell. 24*, 1, 34–58.

Ye, H. and Xu, G. 2003. Fast search in large-scale image database using vector quantization. In *Proceedings of the 2nd International Conference on Image and Video Retrieval*. (July), Urbana, IL, E. M. Bakker, T. S. Huang, M. S. Lew, N. Sebe, and X. Zhou, Eds. Springer-Verlag, London, UK, 477–487.

Yin, P. Y., Bhanu, B., Chang, K. C., and Dong, A. 2005. Integrating relevance feedback techniques for image retrieval using reinforcement learning. *IEEE Trans. Patt. Analy. Machine Intell. 27*, 10, 1536–1551.

Zhou, X. S. and Huang, T. S. 2001. Comparing discriminating transformations and SVM for learning during multimedia retrieval. In *Proceedings of the 9th ACM International Conference on Multimedia*. Ottawa, Canada, 137–146.

Received December 2005; accepted December 2005