Image Retrieval: Ideas, Influences, and Trends of the New Age

RITENDRA DATTA, DHIRAJ JOSHI, JIA LI, and JAMES Z. WANG

The Pennsylvania State University

5

We have witnessed great interest and a wealth of promise in content-based image retrieval as an emerging technology. While the last decade laid foundation to such promise, it also paved the way for a large number of new techniques and systems, got many new people involved, and triggered stronger association of weakly related fields. In this article, we survey almost 300 key theoretical and empirical contributions in the current decade related to image retrieval and automatic image annotation, and in the process discuss the spawning of related subfields. We also discuss significant challenges involved in the adaptation of existing image retrieval techniques to build systems that can be useful in the real world. In retrospect of what has been achieved so far, we also conjecture what the future may hold for image retrieval research.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*; I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms: Algorithms, Documentation, Performance

Additional Key Words and Phrases: Content-based image retrieval, annotation, tagging, modeling, learning

ACM Reference Format:

Datta, R., Joshi, D., Li, J., and Wang, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40, 2, Article 5 (April 2008), 60 pages DOI = 10.1145/1348246.1348248 http://doi.acm.org/10.1145/1348246.1348248

1. INTRODUCTION

What exactly Niels Henrik David Bohr meant when he said "never express yourself more clearly than you are able to think" is anybody's guess. In light of the current discussion, one thought that this well-known quote evokes is that of subtle irony; There are times and situations when we imagine what we desire, but are unable to express this desire in precise wording. Take, for instance, a desire to find the perfect portrait from a collection. Any attempt to express what makes a portrait "perfect" may end up

This material is based on work supported by the National Science Foundation under Grant Numbers 0347148 and 0705210.

J. Z. Wang is also affiliated with the Robotics Institute, Carnegie Meilon University.

Authors' addresses: R. Datta, Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802; email: {datta,djoshi}@cse.psu.edu; D. Joshi, Kodak Research Labs, Rochester, NY 146050; email: dhiraj.joshi@kodak.com; J. Li, Department of Statistics, The Pennsylvania State University, University Park, PA 16802; email: jiali@stat.psu.edu; J. Z. Wang, College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802; email: jwang@ist.psu.edu.

undervaluing the beauty of imagination. In some sense, it may be easier to find such a picture by looking through the collection and making unconscious "matches" with the one drawn by imagination, than to use textual descriptions that fail to capture the very essence of perfection. One way to appreciate the importance of visual interpretation of picture content for indexing and retrieval is this.

Our motivation to organize things is inherent. Over many years we learned that this is a key to progress without the loss of what we already possess. For centuries, text in different languages has been set to order for efficient retrieval, be it manually in the ancient *Bibliotheke*, or automatically as in the modern digital libraries. But when it comes to organizing pictures, man has traditionally outperformed machines for most tasks. One reason which causes this distinction is that text is man's creation, while typical images are a mere replica of what man has seen since birth, concrete descriptions of which are relatively elusive. Add to this the theory that the human vision system has evolved genetically over many centuries. Naturally, the interpretation of what we see is hard to characterize, and even harder to teach a machine. Yet, over the past decade, ambitious attempts have been made to make computers learn to understand, index, and annotate pictures representing a wide range of concepts, with much progress.

Content-based image retrieval (CBIR), as we see it today, is any technology that in principle helps to organize digital picture archives by their visual content. By this definition, anything ranging from an image similarity function to a robust image annotation engine falls under the purview of CBIR. This characterization of CBIR as a field of study places it at a unique juncture within the scientific community. While we witness continued effort in solving the fundamental open problem of robust image understanding, we also see people from different fields, such as, computer vision, machine learning, information retrieval, human-computer interaction, database systems, Web and data mining, information theory, statistics, and psychology contributing and becoming part of the CBIR community [Wang et al. 2006]. Moreover, a lateral bridging of gaps between some of these research communities is being gradually brought about as a by-product of such contributions, the impact of which can potentially go beyond CBIR. Again, what we see today as a few cross-field publications may very well spring into new fields of study in the foreseeable future.

Amidst such marriages of fields, it is important to recognize the shortcomings of CBIR as a real-world technology. One problem with all current approaches is the reliance on visual similarity for judging semantic similarity, which may be problematic due to the *semantic gap* [Smeulders et al. 2000] between low-level content and higher-level concepts. While this intrinsic difficulty in solving the core problem cannot be denied, we believe that the current state-of-the-art in CBIR holds enough promise and maturity to be useful for real-world applications if aggressive attempts are made. For example, Google™ and Yahoo!® are household names today primarily due to the benefits reaped through their use, despite the fact that robust text understanding is still an open problem. Online photo-sharing has become extremely popular with Flickr [Flickr 2002], which hosts hundreds of millions of pictures with diverse content. The video-sharing and distribution forum YouTube has also brought in a new revolution in multimedia usage. Of late, there is renewed interest in the media about potential real-world applications of CBIR and image analysis technologies, as evidenced by publications in *Scintific American* [Mirsky 2006], *Discovery News* [Staedter 2006] and on CNN [2005]. We envision that image retrieval will enjoy a success story in the coming years. We also sense a paradigm shift in the goals of the next-generation CBIR researchers. The need of the hour is to establish how this technology can reach out to the common man in the way text retrieval techniques have. Methods for visual similarity, or even semantic similarity (if ever perfected), will remain techniques for building systems. What the average end-user can hope to gain from using such a system is a different question altogether. For some applications, visual similarity may in fact be more critical than semantic. For others, visual similarity may have little significance. Under what scenarios a typical user feels the need for a CBIR system, what the user sets out to achieve with the system, and how she expects the system to aid in this process are some key questions that need to be answered in order to produce a successful system design. Unfortunately, user studies of this nature have been scarce so far.

Comprehensive surveys exist on the topic of CBIR [Aigrain et al. 1996; Rui et al. 1999; Smeulders et al. 2000; Snoek and Worring 2005], all of which deal primarily with work prior to the year 2000. Surveys also exist on closely related topics such as relevance feedback [Zhou and Huang 2003], high-dimensional indexing of multimedia data [Bohm et al. 2001], face recognition [Zhao et al. 2003] (useful for face-based image retrieval), applications of CBIR to medicine [Muller et al. 2004], and applications to art and cultural imaging [Chen et al. 2005]. Multimedia information retrieval as a broader research area covering video-, audio-, image-, and text analysis has been extensively surveyed [Sebe et al. 2003; Lew et al. 2006]. In our current survey, we restrict the discussion to image-related research only.

One of the reasons for writing this survey is that CBIR, as a field, has grown tremendously after the year 2000 in terms of the people involved and the papers published. Lateral growth has also occurred in terms of the associated research questions addressed, spanning various fields. To validate the hypothesis about growth in publications, we conducted a simple exercise. We searched for publications containing the phrases "Image Retrieval" using Google Scholar [Google Scholar 2004] and the digital libraries of ACM, IEEE, and Springer, within each year from 1995 to 2005. In order to account for: (a) the growth of research in computer science as a whole, and (b) Google's yearly variations in indexing publications, the Google Scholar results were normalized using the publication count for the word "computer" for that year. A plot on another young and fast-growing field within pattern recognition, support vector machines (SVMs), was generated in a similar manner for comparison. The results can be seen in Figure 1. Not surprisingly, the graph indicates similar growth patterns for both fields, although SVM has had faster growth. These trends indicate, given the implicit assumptions, a roughly exponential growth in interest in image retrieval and closely related topics. We also observe particularly strong growth over the last five years, spanning new techniques, support systems, and application domains.

In this article, we comprehensively survey, analyze, and quantify current progress and future prospects of image retrieval. A possible organization of the various facets of image retrieval as a field is shown in Figure 2. Our article follows a similar structure. Note that the treatment is limited to progress mainly in the current decade, and only includes work that involves visual analysis in part or full. For the purpose of completeness, and better readability for the uninitiated, we have introduced key contributions of the earlier years in Section 1.1. Image retrieval purely on the basis of textual metadata, Web link structures, or linguistic tags is excluded. The rest of this article is arranged as follows: For a CBIR system to be useful in the real world, a number of issues need to be taken care of. Hence, the desiderate of real-world image retrieval systems, including various critical aspects of their design, are discussed in Section 2. Some key approaches and techniques of the current decade are presented in detail, in Section 3. Core research in CBIR has given birth to new problems, which we refer to here as CBIR offshoots. These are discussed in Section 4. When distinct solutions to a problem as open-ended as CBIR are proposed, a natural question that arises is how to make a fair comparison among them. In Section 5, we present current directions in the evaluation of image retrieval systems. We conclude in Section 6.

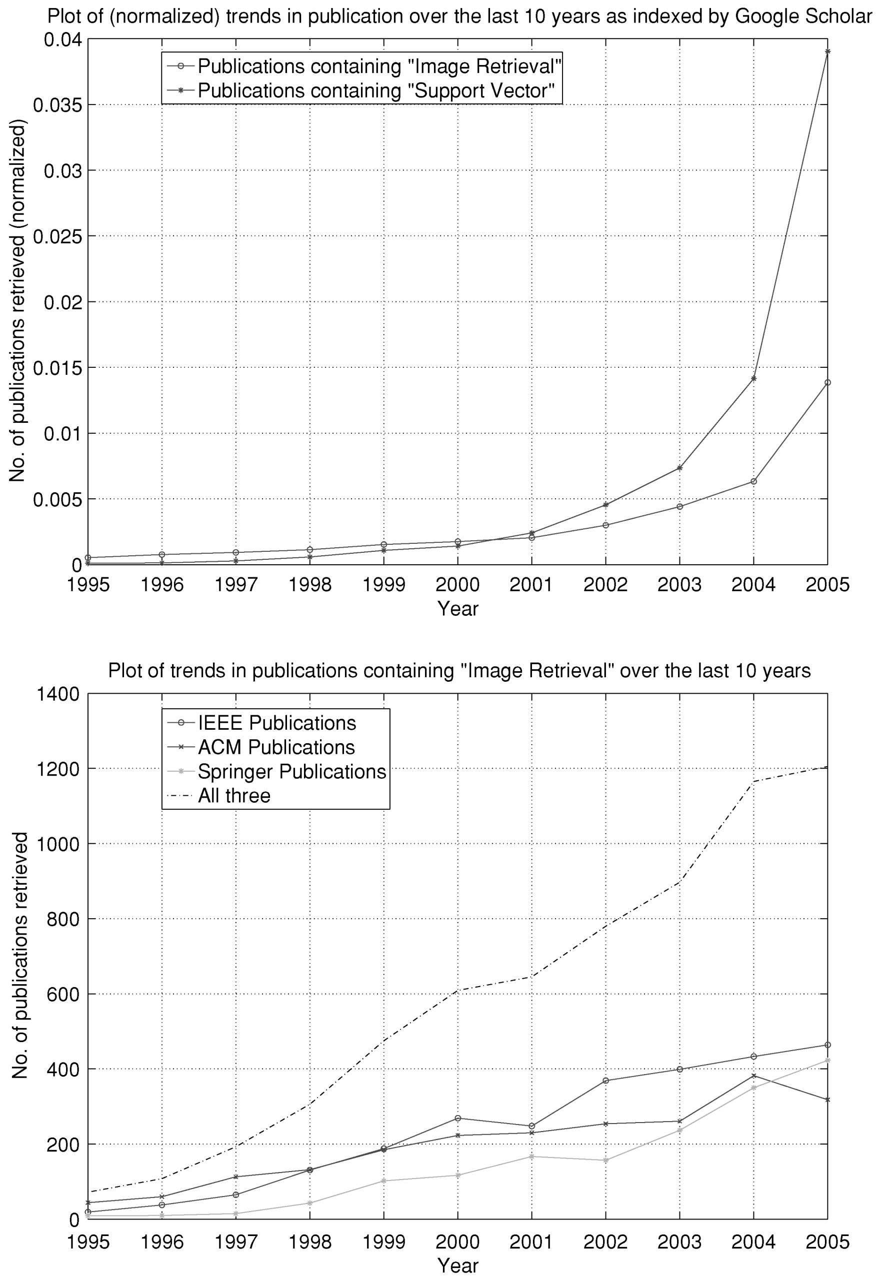


Fig. 1. A study of post-1995 publications in CBIR. Top: Normalized trends in publications containing phrases "image retrieval" and "support vector". Bottom: Publisher, wise break-up of publication count on papers containing "image retrieval".

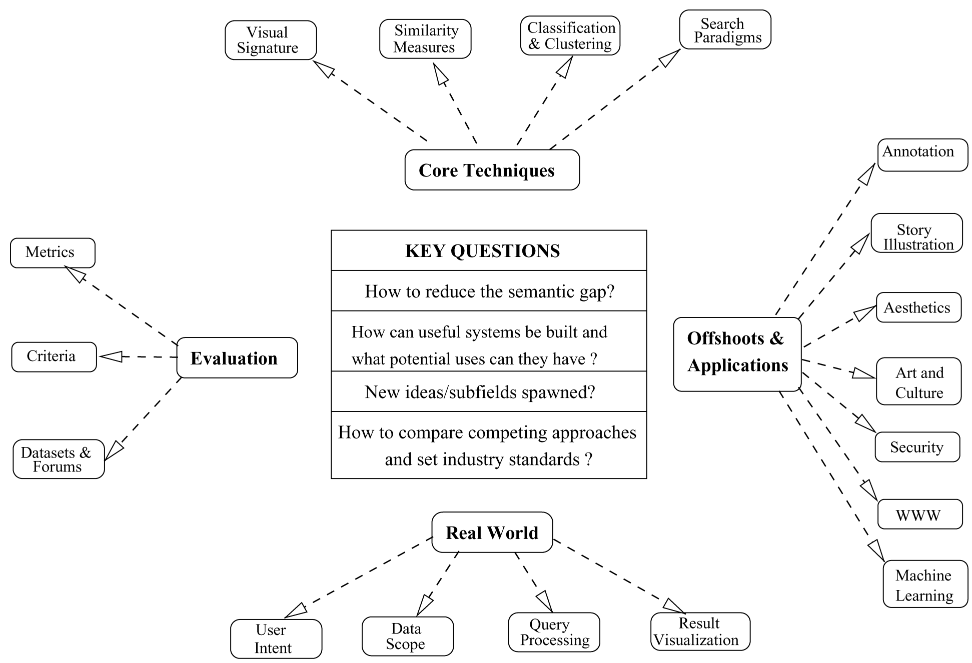


Fig. 2. Our view of the many facets of image retrieval as a field of research. The view is reflected in the structure of this article.

1.1. The Early Years

The years 1994–2000 can be thought of as the initial phase of research and development on image retrieval by content. The progress made during this phase was lucidly summarized at a high level in Smeulders et al. [2000], which has had a clear influence on progress made in the current decade, and will undoubtedly continue to influence future work. Therefore, it is pertinent that we provide a brief summary of the ideas, influences, and trends of the early years (a large part of which originate in that survey) before describing the same for the new age. In order to do so, we first quote the various *gaps* introduced there that define and motivate most of the related problems.

—*Sensory*. The *sensory gap* is the gap between the object in the world and the information in a (computational) description derived from a recording of that scene.
—*Semantic*. The *semantic gap* is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data has for a user in a given situation.

While the former makes recognition from image content challenging due to limitations in recording, the latter brings in the issue of a user's interpretations of pictures and how it is inherently difficult for visual content to capture them. We continue briefly summarizing key contributions of the early years that deal with one or more of these gaps.

In Smeulders et al. [2000], the domains for image search were classified as *narrow* and *broad*, and to-date this remains an extremely important distinction for the purpose of system design. As mentioned, narrow image domains usually have limited variability and better-defined visual characteristics (e.g., aviation-related pictures [Airliners.Net 2005]), which makes content-based image search a tad bit easier to formulate. On the other hand, broad domains tend to have high variability and unpredictability for the same underlying semantic concepts (e.g., Web images), which makes generalization much more challenging. As recently noted in Huijsmans and Sebe [2005], narrow and broad domains pose a problem in image search evaluation as well, and appropriate modifications must be made to standard evaluation metrics for consistency. The survey also lists three broad categories of image search: (1) *search by association*, where there is no clear intent at a picture, but instead the search proceeds by iteratively refined browsing; (2) *aimed search*, where a specific picture is sought; and (3) *category search*, where a single picture representative of a semantic class is sought, for example, to illustrate a paragraph of text, as introduced in Cox et al. [2000]. Also discussed are different kinds of domain knowledge that can help reduce the sensory gap in image search. Notable among them are concepts of syntactic, perceptual, and topological similarity. The overall goal therefore remains to bridge the semantic and sensorial gaps using the available visual features of images and relevant domain knowledge to support the varied search categories, ultimately to satiate the user. We discuss and extend some of these ideas from new perspectives in Section 2.

In the survey, extraction of visual content from images is split into two parts, namely image processing and feature construction. The question to ask here is what features to extract that will help perform meaningful retrieval. In this context, search has been described as a specification of *minimal invariant conditions* that model the user intent, geared at reducing the sensory gap due to accidental distortions, clutter, occlusion, etc. Key contributions in color, texture, and shape abstraction have then been discussed. Among the earliest use of color histograms for image indexing was that in Swain and Ballard [1991]. Subsequently, feature extraction in systems such as QBIC [Flickner et al. 1995], Pictoseek [Gevers and Smeulders 2000], and VisualSEEK [Smith and Chang 1997b] are notable. Innovations in color constancy, that is, the ability to perceive the same color amidst environmental changes, were made by taking specular reflection and shape into consideration [Finlayson 1996]. In Huang et al. [1999] color correlograms were proposed as enhancements to histograms, that take into consideration spatial distribution of colors as well. Gabor filters were successfully used for local shape extraction geared toward matching and retrieval in Manjunath and Ma [1996]. Daubechies' wavelet transforms were used to improve color layout feature extraction in the WBIIS system [Wang et al. 1998]. Viewpoint- and occlusion-invariant local features for image retrieval [Schmid and Mohr 1997] received significant attention as a means to bridge the sensorial gap. Work on local patch-based salient features [Tuytelaars and van Gool 1999] found prominence in areas such as image retrieval and stereo matching. Perceptual grouping of images, important as it is for identifying objects in pictures, is also a very challenging problem. It has been categorized in the survey as strong/weak segmentation (data-driven grouping), partitioning (data-independent grouping, e.g., fixed-image blocks), and sign location (grouping based on a fixed template). Significant progress had been made in field of image segmentation, for example, Zhu and Yuille [1996], where snake- and region growing ideas were combined within a principled framework, and Shi and Malik [2000], where spectral graph partitioning was employed for this purpose. From segments come shape and shape matching needs. In Del Bimbo and Pala [1997], elastic matching of images was successfully applied to sketch-based image retrieval. Image representation by multiscale contour models was studied in Mokhtarian [1995]. The use of graphs to represent spatial relationships between objects, specifically geared toward medical imaging, was explored in Petrakis and Faloutsos [1997]. In Smith and Chang [1997a], 2D-strings [Chang et al. 1987] were employed for characterizing spatial relationships among regions. A method for automatic feature selection was proposed in Swets and Weng [1996]. In Smeulders et al. [2000], the topic of visual content description was concluded with a discussion on the advantages and problems of image segmentation, along with approaches that can avoid strong segmentation while still characterizing image structure well enough for image retrieval. In the current decade, many region-based methods for image retrieval have been proposed that do not depend on strong segmentation. We discuss these and other new innovations in feature extraction in Section 3.1.

Once image features were extracted, the question remained as to how they could be indexed and matched against each other for retrieval. These methods essentially aimed to reduce the semantic gap as much as possible, sometimes reducing the sensorial gap as well in the process. In Smeulders et al. [2000], similarity measures were grouped as feature-based matching (e.g., Swain and Ballard [1991]), object-silhouette-based matching (e.g., Del Bimbo and Pala [1997]), structural feature matching (i.e., hierarchically ordered sets of features, e.g., Wilson and Hancock [1997]), salient feature matching (e.g., geometric hashing Wolfson and Rigoutsos [1997]), matching at the semantic level (e.g., Fagin [1997]), and learning-based approaches for similarity matching (e.g., Wu et al. [2000] and Webe et al. [2000]). Closely tied to the similarity measures are how they emulate the user needs, and, more practically, how they can be modified step-wise with feedback from the user. In this respect, a major advance made in the user interaction technology for image retrieval was relevance feedback (RF). Important early work that introduced RF into the image retrieval domain included Rui et al. [1998], which was implemented in their MARS system [Rui et al. 1997]. Methods for visualization of image query results were explored, for example, in Flickner et al. [1995] and Chang et al. [1997]. Content-based image retrieval systems that gained prominence in this era were, for example, IBM QBIC [Flickner et al. 1995], VIRAGE [Gupta and Jain 1997], and NEC AMORE [Mukherjea et al. 1999] in the commercial domain, and MIT Photobook [Pentland et al. 1994], Columbia VisualSEEK and WebSEEK [Smith and Chang 1997b], UCSB NeTra [Ma and Manjunath 1997], and Stanford WBIIS [Wang et al. 1998] in the academic domain. In Smeulders et al. [2000], practical issues such as system implementation and architecture, as well as their limitations and how to overcome them, the user in the loop, intuitive result visualization, and system evaluation were discussed, and suggestions made. Innovations of the new age based on these suggestions and otherwise are covered extensively in our survey in Sections 2, 3, and 5.

2. IMAGE RETRIEVAL IN THE REAL WORLD

Invention of the digital camera has given the common man the privilege to capture his world in pictures, and conveniently share them with others. One can today generate volumes of images with content as diverse as family get-togethers and national park visits. Low-cost storage and easy Web hosting has fueled the metamorphosis of common man from a passive consumer of photography in the past to a current-day active producer. Today, searchable image data exists with extremely diverse visual and semantic content, spanning geographically disparate locations, and is rapidly growing in size. All these factors have created innumerable possibilities and hence considerations for real-world image search system designers.

As far as technological advances are concerned, growth in content-based image retrieval has been unquestionably rapid. In recent years, there has been significant effort put into understanding the real world implications, applications, and constraints of the technology. Yet, real-world application of the technology is currently limited. We devote this section to understanding image retrieval in the real world and discuss user expectations, system constraints and requirements, and the research effort to make image retrieval a reality in the not-too-distant future.

Designing an omnipotent real-world image search engine capable of serving all categories of users requires understanding and characterizing user-system interaction and image search, from both user and system points-of-view. In Figure 3, we propose

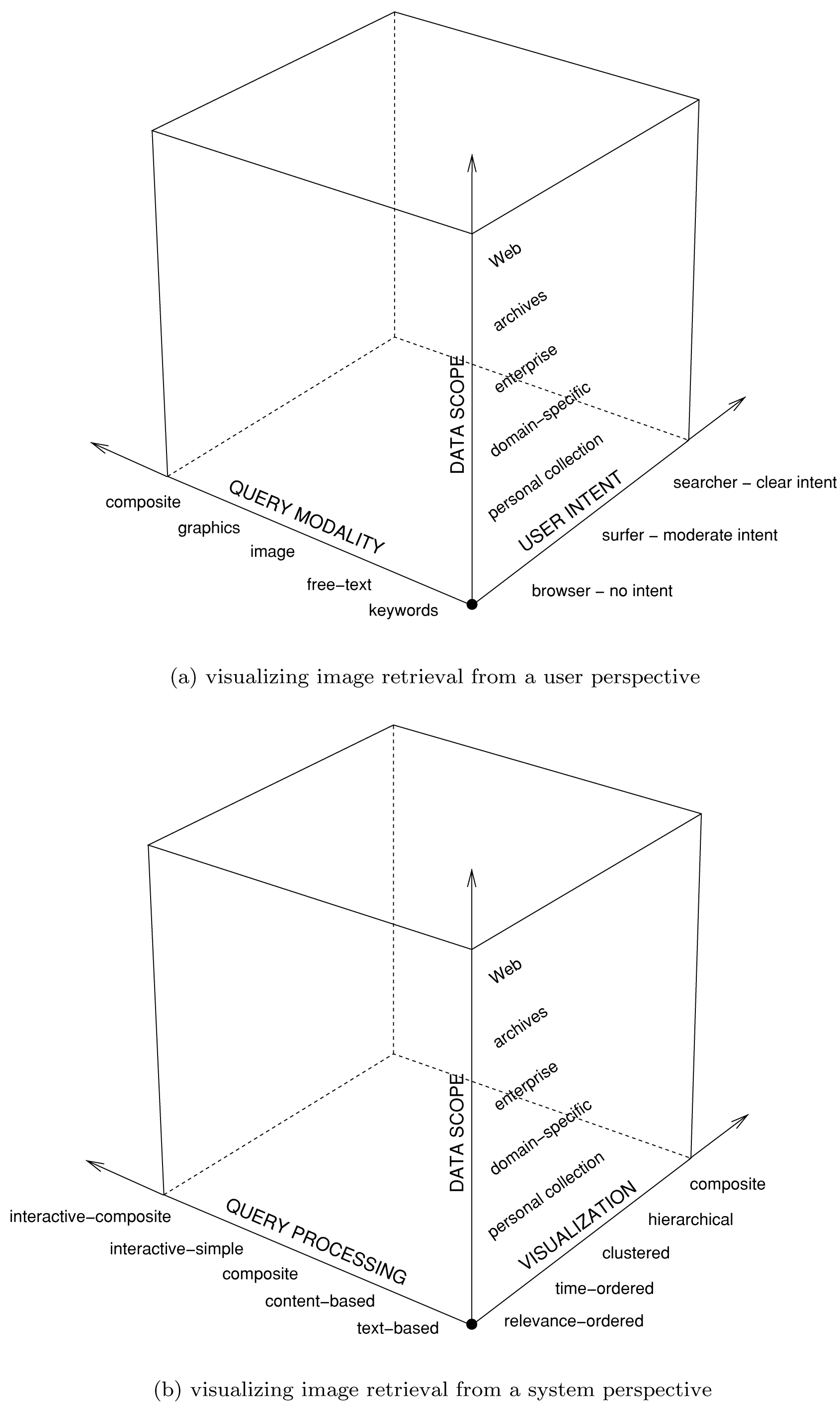


Fig. 3. Our views of image retrieval from user and system perspectives.

one such dual characterization, and attempt to represent all known possibilities of interaction and search. From a *user perspective*, embarking on an image search, journey involves considering and making decisions on the following fronts: (1) clarity of the user about what she wants, (2) where she wants to search, and (3) the form in which the user has her query. In an alternative view from an image retrieval *system perspective*, a search translates to making arrangements as per the following factors: (1) how does the user wish the results to be presented, (2) where does the user desire to search, and (3) what is the nature of user input/interaction. These factors, with their respective possibilities, form our axes for Figure 3. In the proposed user and system spaces, real-world image search instances can be considered as isolated points or point clouds, and search sessions can consist of trajectories while search engines can be thought of as surfaces. The intention of drawing cubes versus free 3D Cartesian spaces is to emphasize that the possibilities are indeed bounded by the size of the Web, the nature of user, and ways of user-system interaction. We believe that the proposed characterization will be useful for designing context-dependent search environments for real-world image retrieval systems.

2.1. User Intent

We augment the search-type-based classification proposed in Smeulders et al. [2000] with a user-intent-based classification. When users search for pictures, their intent or clarity about what they desire may vary. We believe that clarity of intent plays a key role in a user's expectation from a search system and the nature of her interaction. It can also act as a guideline for system design. We broadly characterize a user by clarity of her intent as follows.

—*Browser*. This is a user browsing for pictures with no clear end-goal. A browser's session would consist of a series of unrelated searches. A typical browser would jump across multiple topics during the course of a search session. Her queries would be incoherent and diverse in topic.

—*Surfer*. A surfer is a user surfing with moderate clarity of an end-goal. A surfer's actions may be somewhat exploratory in the beginning, with the difference that subsequent searches are expected to increase the surfer's clarity of what she wants from the system.

—*Searcher*. This is a user who is very clear about what she is searching for in the system. A searcher's session would typically be short, with coherent searches leading to an end-result.

A typical browser values ease of use and manipulation. A browser usually has plenty of time at hand and expects surprises and random search hints to elongate her session (e.g., picture of the day, week, etc.). On the other hand, a surfer would value a search environment which facilitates clarity of her goal. A surfer planning a holiday would value a hint such as "pictures of most popular destinations". At the other extreme, the searcher views an image retrieval system from a core utilitarian perspective. Completeness of results and clarity of representation would usually be the most important factors. The impact of real-world usage from the user viewpoint has not been extensively studied. One of the few studies categorizes users as *experts* and *novices* and studies their interaction patterns with respect to a video library [Christel and Conescu 2005]. In Armitage and Enser [1997], an analysis of user needs for visual information retrieval was conducted. In the cited work, a categorization schema for user queries was proposed, with a potential to be embedded in the visual information retrieval system.

Discussion. In the end, all that matters to an end-user is her interaction with the system, and the corresponding response. The importance of building human-centered multimedia systems has been expressed lately [Jaimes et al. 2006]. In order to gain wide acceptance, image retrieval systems need to acquire a human-centered perspective as well.

2.2. Data Scope

Understanding the nature and scope of image data plays a key role in the complexity of image search system design. Factors such as the diversity of user-base and expected user traffic for a search system also largely influence the design. Along this dimension, we classify search data into the following categories.

—*Personal Collection.* This consists of a largely homogeneous collection generally small in size, accessible primarily to its owner, and usually stored on a local storage media.

—*Domain-Specific Collection.* This is a homogeneous collection providing access to controlled users with very specific objectives. The collection may be large and hosted on distributed storage, depending upon the domain. Examples of such a collection are biomedical and satellite image databases.

—*Enterprise Collection.* We define this as a heterogeneous collection of pictures accessible to users within an organization's intranet. Pictures may be stored in many different locations. Access may be uniform or nonuniform, depending upon the Intranet design.

—*Archives.* These are usually of historical interest and contain large volumes of structured or semi-structured homogeneous data pertaining to specific topics. Archives may be accessible to most people on the Internet, with some control of usage. Data is usually stored in multiple disks or large disk arrays.

—*Web.* World Wide Web pictures are accessible to practically everyone with an Internet connection. Current WWW image search engines such as Google and Yahoo! images have a key crawler component which regularly updates their local database to reflect on the dynamic nature of the Web. Image collection is semi-structured, nonhomogeneous, and massive in volume, and is usually stored in large disk arrays.

An image retrieval system designed to serve a personal collection should focus on features such as personalization, flexibility of browsing, and display methodology. For example, Google's Picasa system [Picasa 2004] provides a chronological display of images taking a user on a journey down memory lane. Domain-specific collections may impose specific standards for presentation of results. Searching an archive for content discovery could involve long user search sessions. Good visualization and a rich query support system should be the design goals. A system designed for the Web should be able to support massive user traffic. One way to supplement software approaches for this purpose is to provide hardware support to the system architecture. Unfortunately, very little has been explored in this direction, partly due to the lack of agreed-upon indexing and retrieval methods. The notable few applications include an FPGA implementation of a color-histogram-based image retrieval system [Kotoulas and Andreadis 2003], an FPGA implementation for subimage retrieval within an image database [Nakano and Takamichi 2003], and a method for efficient retrieval in a network of imaging devices [Woodrow and Heinzelman 2002].

Discussion. Regardless of the nature of the collection, as the expected user-base grows, factors such as concurrent query support, efficient caching, and parallel and distributed processing of requests become critical. For future real-world image retrieval systems, both software and hardware approaches to address these issues are essential.

More realistically, dedicated specialized servers, optimized memory and storage support, and highly parallelizable image search algorithms to exploit cluster computing powers are where the future of large-scale image search hardware support lies.

2.3. Query Modalities and Processing

In the realm of image retrieval, an important parameter to measure user-system interaction level is the complexity of queries supported by the system. From a user perspective, this translates to the different modalities she can use to query a system. We describe next the various querying modalities, their characteristics, and the system support required thereof.

—*Keywords*. This is a search in which the user poses a simple query in the form of a word or bigram. This is currently the most popular way to search images, for example, the Google and Yahoo! image search engines.

—*Free-Text*. This is where the user frames a complex phrase, sentence, question, or story about what she desires from the system.

—*Image*. Here, the user wishes to search for an image similar to a query image. Using an example image is perhaps the most representative way of querying a CBIR system in the absence of reliable metadata.

—*Graphics*. This consists of a hand-drawn or computer-generated picture, or graphics could be presented as query.

—*Composite*. These are methods that involve using one or more of the aforesaid modalities for querying a system. This also covers interactive querying such as in relevance feedback systems.

The aforementioned query modalities require different processing methods and/or support for user interaction. The processing becomes more complex when visual queries and/or user interactions are involved. We next broadly characterize query processing from a system perspective.

—*Text-Based*. Text-based query processing usually boils down to performing one or more simple keyword-based searches and then retrieving matching pictures. Processing a free text could involve parsing, processing, and understanding the query as a whole. Some form of natural language processing may also be involved.

—*Content-Based*. Content-based query processing lies at the heart of all CBIR systems. Processing of query (image or graphics) involves extraction of visual features and/or segmentation and search in the visual feature space for similar images. An appropriate feature representation and a similarity measure to rank pictures, given a query, are essential here. These will be discussed in detail in Section 3.

—*Composite*. Composite processing may involve both content- and text-based processing in varying proportions. An example of a system which supports such processing is the story picturing engine [Joshi et al. 2006b].

—*Interactive-Simple*. User interaction using a single modality needs to be supported by a system. An example is a relevance-feedback-based image retrieval system.

—*Interactive-Composite*. The user may interact using more than one modality (e.g., text and images). This is perhaps the most advanced form of query processing that is required to be performed by an image retrieval system.

Processing text-based queries involves keyword matching using simple set-theoretic operations, and therefore a response can be generated very quickly. However, in very large systems working with millions of pictures and keywords, efficient indexing methods may be required. Indexing of text has been studied in database research for decades now. Efficient indexing is critical to the building and functioning of very large text-based databases and search engines. Research on efficient ways to index images by content has been largely overshadowed by research on efficient visual representation and similarity measures. Most of the methods used for visual indexing are adopted from text-indexing research. In Petrakis et al. [2002], R-trees are used for indexing images represented as attributed relational graphs (ARGs). Retrieval of images using wavelet coefficients as image representations and R^*-trees for indexing has been studied in Natsev et al. [2004]. Visual content matching using graph-based image representation and an efficient metric indexing algorithm has been proposed in Berretti et al. [2001]. More details of techniques for content-based indexing of pictures can be found in Marsicoi et al. [1997] and Del Bimbo [1999].

Composite querying methods provide the users with more flexibility for expressing themselves. Some recent innovations in querying include sketch-based retrieval of color images [Chalechale et al. 2005]. Querying using 3D models [Assfalg et al. 2002] has been motivated by the fact that 2D image queries are unable to capture the spatial arrangement of objects within the image. In another interesting work, a multimodal system involving hand gestures and speech for querying and relevance feedback was presented in Kaster et al. [2003]. Certain new interaction-based querying paradigms which statistically model the user's interest [Fang et al. 2005], or help the user refine her queries by providing cues and hints [Jaimes et al. 2004; Nagamine et al. 2004], have been explored for image retrieval.

Use of mobile devices has become widespread lately. Mobile users have limited querying capabilities due to inherent scrolling and typing constraints. Relevance feedback has been explored for quickly narrowing down search to such user needs. However, mobile users can be expected to provide only limited feedback. Hence, it becomes necessary to design intelligent feedback methods to cater to users with small displays. The performance of different relevance feedback algorithms for small devices has been studied and compared in Vinay et al. [2005, 2004]. In the cited work, a tree-structured representation for all possible user-system actions was used to determine an upper bound on the performance gains that such systems can achieve.

Discussion. A prerequisite for supporting text-based query processing is the presence of reliable metadata with pictures. However, pictures rarely come with reliable human tags. In recent years, there has been effort put into building interactive, public-domain games for large-scale collection of high-level manual annotations. One such game (the ESP game) has become very popular and has helped accumulate human annotations for about a hundred thousand pictures [von Ahn and Dabbish 2004]. Collection of manual tags for pictures has the dual advantage of: (1) facilitating text-based querying, and (2) building reliable training datasets for content-based analysis and automatic annotation algorithms. As explored in Datta et al. [2007], it is possible to effectively bridge the paradigms of keyword- and content-based search through a unified framework to provide the user the flexibility of both, without losing out on the search scope.

2.4. Visualization

Presentation of search results is perhaps one of the most important factors in the acceptance and popularity of an image retrieval system. We characterize common visualization schemes for image search as follows.

—*Relevance-Ordered*. The most popular way to present search results is relevance-ordered, as adopted by Google and Yahoo! for their image search engines. Results are ordered by some numeric measure of relevance to the query.

—*Time-Ordered*. In time-ordered image search, pictures are shown in a chronological ordering rather than by relevance. Google's Picasa system [Picasa 2004] for personal collections provides an option to visualize a chronological timeline using pictures.

—*Clustered*. Clustering of images by their metadata or visual content has been an active research topic for several years (discussed in Section 3). Clustering of search results, besides being an intuitive and desirable form of presentation, has also been used to improve retrieval performance [Chen et al. 2005].

—*Hierarchical*. If metadata associated with images can be arranged in tree order (e.g., WordNet topical hierarchies [Miller 1995]), it can be a very useful aid in visualization. Hierarchical visualization of search results is desirable for archives, especially for educational purposes.

—*Composite*. Combining consists of mixing two or more of the preceding forms of visualization scheme, and is used especially for personalized systems. Hierarchical clustering and visualization of concept graphs are examples of composite visualizations.

In order to design interfaces for image retrieval systems, it helps to understand factors like how people manage their digital photographs [Rodden and Wood 2003] or frame their queries for visual art images Cunningham et al. [2004]. In Rodden et al. [2001], user studies on various ways of arranging images for browsing purposes are conducted, and the observation is that both visual-feature-based and concept-based arrangements have their own merits and demerits. Thinking beyond the typical grid-based arrangement of top matching images, spiral and concentric visualization of retrieval results have been explored in Torres et al. [2003]. For personal images, innovative arrangements of query results based on visual content, time-stamps, and efficient use of screen space add new dimensions to the browsing experience [Huynh et al. 2005].

Portable devices such as personal digital assistants (PDAs) and vehicle communication and control systems are becoming very popular as client-side systems for querying and accessing remote multimedia databases. A portable-device user is often constrained in the way she can formulate her query and interact with a remote image server. There are inherent scrolling and browsing constraints which can constrict user feedback. Moreover, there are bandwidth limitations which need to be taken into consideration when designing retrieval systems for such devices. Some additional factors which become important here are size and color depth of display. Personalization of search for small displays by modeling interaction from the gathered usage data has been proposed in Bertini et al. [2005]. An image attention model for adapting images based on user attention for small displays has been proposed in Chen et al. [2003]. Efficient ways of browsing large images interactively, such as those encountered in pathology or remote sensing, using small displays over a communication channel are discussed in Li and Sun [2003]. User-log-based approaches to smarter ways of image browsing on mobile devices have been proposed in Xie et al. [2005].

Image transcoding techniques, which aim at adapting multimedia (image and video) content to the capabilities of the client device, have been studied extensively in the last several years [Shanableh and Ghanbari 2000; Vetro et al. 2003; Bertini et al. 2003; Cucchiara et al. 2003]. A class of methods known as semantic transcoding aims at designing intelligent transcoding systems which can adapt semantically to user requirements [Bertini et al. 2003; Cucchiara et al. 2003]. For achieving this, classes of relevance are constructed and transcoding systems are programmed differently for different classes.

Discussion. Study of organizations which maintain image management and retrieval systems has provided useful insights into system design, querying, and visualization. In Tope and Enser [2000], case studies on the design and implementation of many different electronic retrieval systems have been reported. The final verdict of acceptance/rejection for any visualization scheme comes from end-users. While simple, intuitive interfaces such as grid-based displays have become acceptable to most search engine users, advanced visualization techniques could still be in the making. It becomes critical for visualization designers to ensure that the added complexity does not become an overkill.

2.5. Real-World Image Retrieval Systems

Not many image retrieval systems are deployed for public usage, save for Google Image Search or Yahoo! Image Search (which are based primarily on surrounding metadata such as filenames and HTML text). Recently, a public-domain search engine called *Riya* (see Figure 4) has been developed, which incorporates image retrieval and face recognition for searching pictures of people and products on the Web. It is also interesting to note that CBIR technology is being applied to domains as diverse as family album management, botany, astronomy, mineralogy, and remote sensing [Zhang et al. 2003; Wang et al. 2002; Csillaghy et al. 2000; Painter et al. 2003; Schroder et al. 2000]. A publicly available similarity search tool [Wang et al. 2001] is being used for an online database of over 800, 000 airline-related images [Airliners.Net 2005; Slashdot 2005] (again see Figure 4), the integration of similarity search functionality to a large collection of art and cultural images [GlobalMemoryNet 2006], and the incorporation of image similarity to a massive picture archive [Terragalleria 2001] of the renowned travel photographer Q.-T. Luong.

Automatic Linguistic Indexing of Pictures—Real-Time (ALIPR), an automatic image annotation system [Li and Wang 2006a; 2008], has been recently made public for people to try to have their pictures annotated. As mentioned earlier, presence of reliable tags with pictures is necessary for text-based image retrieval. As part of the ALIPR search engine, an effort to automatically validate computer generated tags with human-given annotation is being used in an attempt to build a very large collection of searchable images (see Figure 5). Another work-in-progress is a Web image search system [Joshi et al. 2006a] that exploits visual features and textual metadata, using state-of-the-art algorithms, for a comprehensive search experience.

Discussion. Image analysis and retrieval systems have received widespread public and media interest of late [Mirsky 2006; Staedter 2006; CNN 2005]. It is reasonable to hope that in the near future, the technology will diversify to many other domains. We believe that the future of real-world image retrieval lies in exploiting both text- and content-based search technologies. While the former is considered more reliable from a user viewpoint, there is immense potential in combining the two to build robust image search engines that would make the "hidden" part of the Web images accessible. This endeavor will hopefully be actualized in the years to come.

3. IMAGE RETRIEVAL TECHNIQUES: ADDRESSING THE CORE PROBLEM

Despite the effort made in the early years of image retrieval research (Section 1.1), we do not yet have a universally acceptable algorithmic means of characterizing human vision, more specifically in the context of interpreting images. Hence, it is not surprising to see continued effort in this direction, either building up on prior work or exploring novel directions. Considerations for successful deployment of CBIR in the real world are reflected by the research focus in this area.

By the nature of its task, the CBIR technology boils down to two intrinsic problems: (a) how to mathematically describe an image, and (b) how to assess the similarity

Fig. 4. Real-world use of content-based image retrieval using color, texture, and shape matching. Top: `http://airliners.net` is a photo-sharing community with more than a million airplane-related pictures. Bottom: `http://riya.com` is a collection of several million pictures.

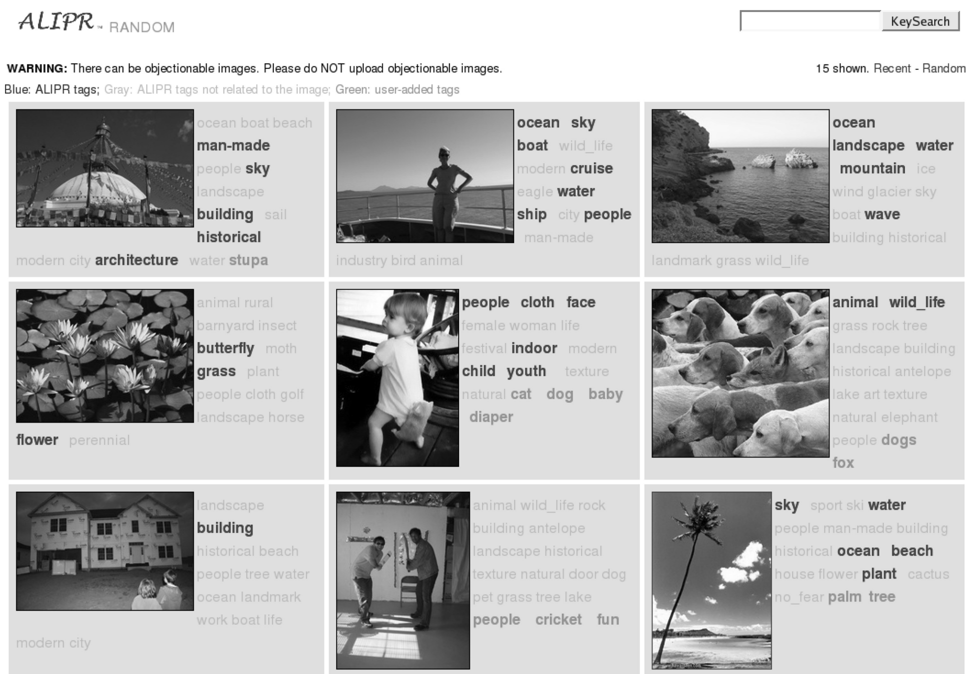


Fig. 5. Real-world use of *automatic image annotation* from `http://alipr.com`. The screenshot shows a random set of uploaded pictures and with annotations given both by ALIPR (shown in dark and light gray) and by users (shown in medium gray and printed at the end of each block).

between a pair of images based on their abstracted descriptions. The first issue arises because the original representation of an image which is an array of pixel values, corresponds poorly to our visual response, let alone semantic understanding of the image. We refer to the mathematical description of an image, for retrieval purposes, as its *signature*. From the design perspective, the extraction of signatures and the calculation of image similarity cannot be cleanly separated. The formulation of signatures determines to a large extent the realm for definitions of similarity measures. On the other hand, intuitions are often the early motivating factors for designing similarity measures in a certain way, which in turn puts requirements on the construction of signatures.

In comparison with earlier, pre-2000 work in CBIR, a remarkable difference of recent years has been the increased diversity of image signatures. Advances have been made in both the derivation of new features (e.g., shape) and the construction of signatures based on these features, with the latter type of progress being more pronounced. The richness in the mathematical formulation of signatures grows alongside the invention of new methods for measuring similarity. In the rest of this section, we will first address the extraction of image signatures, and then the methods for computing image similarity based on the signatures. In terms of methodology development, a strong trend which has emerged in recent years is the employment of statistical and machine learning techniques in various aspects of the CBIR technology. Automatic learning, mainly clustering and classification, is used to form either fixed or adaptive signatures, to tune similarity measures, and even to serve as the technical core of certain searching schemes, for example, relevance feedback. We thus not only discuss the influence of learning while addressing fundamental issues of retrieval, but also devote a subsection on clustering and classification in the context of CBIR. Finally, we review different

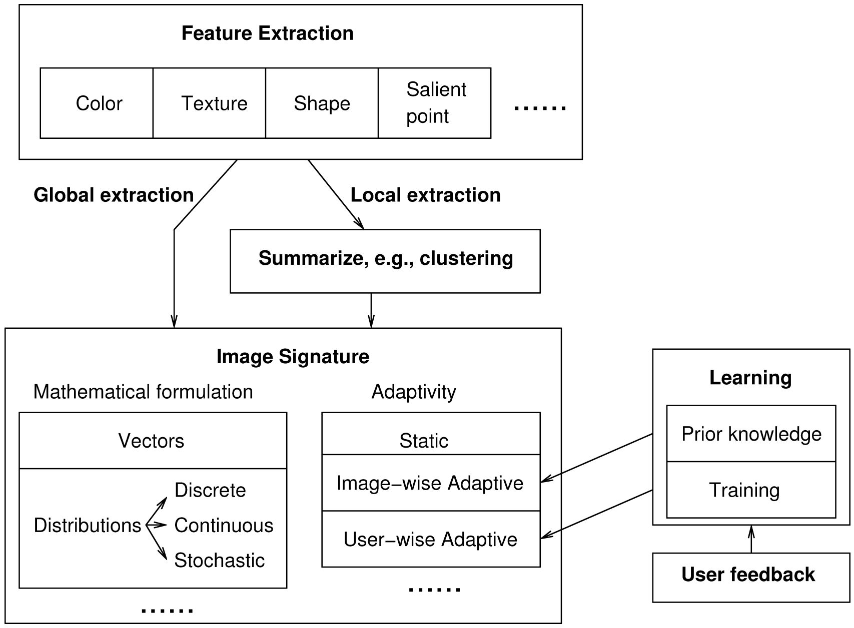


Fig. 6. An overview of image signature formulation.

paradigms of searching with emphasis on relevance feedback. An actively pursued direction in image retrieval is to engage humans in the searching process, that is, to include a *human in the loop*. Although in the very early days of CBIR, several systems were designed with detailed user-preference specifications, the philosophy of engaging users in recent work has evolved toward more interactive and iterative schemes by leveraging learning techniques. As a result, the overhead for a user, in specifying what she is looking for at the beginning of a search, is much reduced.

3.1. Extraction of Visual Signature

Most CBIR systems perform feature extraction as a preprocessing step. Once obtained, visual features act as inputs to subsequent image analysis tasks, such as similarity estimation, concept detection, or annotation. Figure 6 illustrates the procedure of generating image signatures and the main research problems involved. Following the order typical in feature extraction and processing, we present in the following the prominent recent innovations in visual signature extraction. The current decade has seen great interest in region-based visual signatures, for which segmentation is the quintessential first step. While we begin the discussion with recent progress in image segmentation, we will see in the subsequent section that there is significant interest in segmentation-free techniques to feature extraction and signature construction.

Image Segmentation. To acquire a region-based signature, a key step is to segment images. Reliable segmentation is especially critical for characterizing shapes within images, without which the shape estimates are largely meaningless. We described earlier a widely used segmentation approach based on k-means clustering. This basic approach enjoys a speed advantage, but is not as refined as some recently developed methods. One of the most important new advances in segmentation employs the normalized cuts criterion [Shi and Malik 2000]. The problem of image segmentation is mapped to a weighted graph partitioning problem, where the vertex set of the graph is composed of image pixels and the edge weights represent some perceptual similarity between pixel pairs. The normalized cut segmentation method in Shi and Malik [2000] is also extended to textured image segmentation by using cues of contour and texture differences [Malik et al. 2001], and to incorporate known partial grouping priors by solving a constrained optimization problem [Yu and Shi 2004]. The latter has potential for incorporating real-world application-specific priors, such as the location and size cues of organs in pathological images.

Searching of medical image collections has been an increasingly important research problem of late, due to the high-throughput, high-resolution, and high-dimensional imaging modalities introduced. In this domain, 3D brain magnetic resonance (MR) images have been segmented using hidden Markov random fields and the expectation-maximization (EM) algorithm [Zhang et al. 2001], and the spectral clustering approach has found some success in segmenting vertebral bodies from sagittal MR images [Carballido-Gamio et al. 2004]. Among other recent approaches proposed are segmentation based on the mean shift procedure [Comaniciu and Meer 2002], multiresolution segmentation of low-depth-of-field images [Wang et al. 2001], a Bayesian-framework-based segmentation involving the Markov chain Monte Carlo technique [Tu and Zhu 2002], and an EM-algorithm-based segmentation using a Gaussian mixture model [Carson et al. 2002], forming *blobs* suitable for image querying and retrieval. A sequential segmentation approach that starts with texture features and refines segmentation using color features is explored in Chen et al. [2002]. An unsupervised approach for segmentation of images containing homogeneous color/texture regions has been proposed in Deng and Manjunath [2001].

While there is no denying that achieving good segmentation is a major step toward image understanding, some issues plaguing current techniques are computational complexity, reliability of good segmentation, and acceptable segmentation quality assessment methods. In the case of image retrieval, some strategies for getting around this problem have been to reduce dependence on reliable segmentation [Carson et al. 2002], to involve every generated segment of an image in the matching process to obtain *soft* similarity measures [Wang et al. 2001], or to characterize spatial arrangement of color and texture using block-based 2D multiresolution hidden Markov models (MHMMs) [Li et al. 2000; Li and Wang 2003]. Another alternative is to use perceptual grouping principles to hierarchically extract image structures [Iqbal and Aggarwal 2002]. In Datta et al. [2007], probabilistic modeling of class-wise color segment interactions has been employed for the purpose of image categorization and retrieval, to reduce sensitivity to segmentation.

Major Types of Features. A feature is defined to capture a certain visual property of an image, either globally for the entire image or locally for a small group of pixels. The most commonly used features include those reflecting color, texture, shape, and salient points in an image, each of which will be discussed shortly. In global extraction, features are computed to capture the overall characteristics of an image. For instance, in a color layout approach, an image is divided into a small number of subimages and the average color components (e.g., red, green, and blue intensities) are computed for every subimage. The overall image is thus represented by a vector of color components where a particular dimension of the vector corresponds to a certain subimage location. The advantage of global extraction is its high speed for both extracting features and computing similarity. However, as evidenced by the rare use of color layout in recent work, global features are often too rigid to represent an image. Specifically, they can be oversensitive to location and hence fail to identify important visual characteristics. To increase the robustness to spatial transformation, the second approach to form signatures is by local extraction and an extra step of feature summarization.

In local feature extraction, a set of features are computed for every pixel using its neighborhood (e.g., average color values across a small block centered around the pixel). To reduce computation, an image may be divided into small, nonoverlapping blocks, and features are computed individually for every block. The features are still local because of the small block size, but the amount of computation is only a fraction of that for obtaining features around every pixel. Let the feature vectors extracted at block or pixel location (i, j) be $x_{i,j}$, $1 \leq i \leq m$, $1 \leq j \leq n$, where the image size $m \times n$ can vary. To achieve a global description of an image, various ways of summarizing the dataset $\{x_{i,j}, 1 \leq i \leq m, 1 \leq j \leq n\}$ have been explored, leading to different types of signature. A common theme of summarization is to derive a distribution for $x_{i,j}$ based on the dataset.

Exploration of color features was active in nascent CBIR, with emphasis on exploiting color spaces (e.g., LUV) that seem to coincide better with human vision than the basic RGB color space. In recent years, research on color features has focused more on the summarization of colors in an image, that is, the construction of signatures out of colors. A set of color and texture descriptors tested for inclusion in the MPEG-7 standard, and well suited to natural images and video, is described in Manjunath et al. [2001]. These include histogram-based descriptors, spatial color descriptors, and texture descriptors suited for retrieval.

Texture features are intended to capture the granularity and repetitive patterns of surfaces within in a picture. For instance, grassland, brick walls, teddy bears, and flower petals differ in texture, by smoothness as well as patterns. Their role in domain-specific image retrieval, such as in aerial imagery and medical imaging, is particularly vital due to their close relation to the underlying semantics in these cases. Texture features have long been studied in image processing, computer vision, and computer graphics [Haralick 1979], such as multiorientation filter banks [Malik and Perona 1990] and wavelet transforms [Unser 1995]. In image processing, a popular way to form texture features is by using the coefficients of a certain transform on the original pixel values, or, more sophisticatedly, by statistics computed from these coefficients. Examples of texture features using the *wavelet transform* and the discrete cosine transform can be found in Do and Vetterli [2002] and Li et al. [2000]. In computer vision and graphics, advances have been made in fields such as texture synthesis, where Markov statistical descriptors based on pairs of wavelet coefficients at adjacent location/orientation/scale in the images are used [Portilla and Simoncelli 2000]. Among the earliest work on the use of texture features for image retrieval is Manjunath and Ma [1996]. Texture descriptors apt for inclusion in the MPEG-7 were broadly discussed in Manjunath et al. [2001]. Such descriptors encode significant, general visual characteristics into standard numerical formats that can used for various higher-level tasks. A *thesaurus* for texture, geared toward aerial image retrieval, has been proposed in Ma and Manjunath [1998]. The texture extraction part of this thesaurus building process involves the application of a bank of Gabor filters [Jain and Farrokhnia 1990] to the images, to encode statistics of the filtered outputs as texture features. Advances in textured region descriptors have been made, such as affine- and photometric-transformation-invariant features that are also robust to the shape of the region in question [Schaffalitzky and Zisserman 2001]. While the target application is the more traditional stereo matching, it has been shown to have potential for textured image matching and for segmentation as well. Advances in affine-invariant texture feature extraction, designed for texture recognition, have been made in Mikolajczyk and Schmid [2004], with the use of interest point detection for sparsity. Texture features at a point in the image are meaningful only as a function of its neighborhood, and the (effective) size of this neighborhood can be thought of as a *scale* at which these features are computed. Because a choice of scale is critical to the meaningfulness of such features, it has been explored as an automatic scale selection problem in Carson et al. [2002], specifically to aid image retrieval.

Shape is a key attribute of segmented image regions, and its efficient and robust representation plays an important role in retrieval. Synonymous with shape representation is the way in which such representations are matched with each other. Here we discuss both shape representations and the particular forms of shape similarities used in each case. In general, over the years we have seen a shift from global shape representations (e.g., in Flickner et al. [1995]) to more local descriptors (e.g., in Mehrotra and Gary [1995], Berretti et al. [2000], and Petrakis et al. [2002]) due to the typical modeling limitations. Representation of shape using discrete curve evolution to simplify contours is discussed in Latecki and Lakamper [2000]. This contour simplification helps to remove noisy and irrelevant shape features from consideration. A new shape descriptor for similarity matching, referred to as *shape context*, is proposed, which is fairly compact yet robust to a number of geometric transformations [Belongie et al. 2002]. In Berretti et al. [2000], curves are represented by a set of segments, or *tokens*, whose feature representations (curvature and orientation) are arranged into a metric tree [Ciaccia et al. 1997] for efficient shape matching and shape-based image retrieval. A dynamic programming (DP) approach to shape matching is proposed in Petrakis et al. [2002], where shapes are approximated as sequences of concave and convex segments. One problem with this approach is that the computation of Fourier descriptors and moments is slow, although precomputation may help produce real-time results. Continuing with Fourier descriptors, exploitation of both the amplitude and phase, as well as the use of dynamic time warping (DTW) distance instead of Euclidean distance, is shown an accurate shape matching technique in Bartolini et al. [2005]. The rotational and starting point invariance otherwise obtained by discarding the phase information is maintained here by adding compensation terms to the original phase, thus allowing its exploitation for better discrimination.

Closely associated with these methods are approaches that model spatial relations among local image entities for retrieval. Much of the approaches to spatial modeling and matching have been influenced by earlier work on *iconic indexing* Chang et al. [1987, 1988] based on the theory of symbolic projections. Here, images are represented based on orthogonal projections of constituent entities, by encoding the corresponding bidirectional arrangement on the two axes as a *2D string* of entities and relationships. This way, image matching is effectively converted from a spatial matching problem to a 1D matching one. Many variants of the 2D-string model have been proposed since. In recent years, extensions such as 2D-Be-string [Wang 2003] have been proposed, where the symbolic encoding has been extended to represent entity locations more precisely, thus avoiding cutting the entities along their bounding rectangles, for improved complexity. Another work on iconic indexing can be found in Petraglia et al. [2001], where a symbolic representation of real images, termed as a *virtual image*, is proposed, consisting of entities and the binary spatial relations among them. Compared to traditional iconic representations and their variants, this approach allows more explicit scene representation and more efficient retrieval, once again without requiring the entities to be cut. In Berretti et al. [2003], a novel alternative to the previously discussed class of spatial models, *weighted walkthroughs*, is proposed. This representation allows quantitative comparison (which is challenging for purely Boolean relationships) of entities, by incorporating the spatial relationships among each pair of pixels from the two entities. These quantitative relations allow images to be represented by attributed relational graphs (ARGs), which essentially makes the retrieval problem one of graph comparison, resulting in improved retrieval performance over other representations. This idea has been extended to spatial modeling of 3D objects in Berretti and Del Bimbo [2006].

Other image models that capture spatial arrangements between local features, such as interest points, are discussed in the following paragraph.

Features based on local invariants such as *corner points* or *interest points*, traditionally used for stereo matching, are being used in image retrieval as well. Scale- and affine-invariant interest points that can deal with significant affine transformations and illumination changes have been shown effective for image retrieval [Mikolajczyk and Schmid 2004]. Along similar lines, wavelet-based *salient points* have been used for retrieval [Tian et al. 2001]. In more recent work, the earth mover's distance [Rubner et al. 2000] has been used for matching locally invariant features in Grauman and Darrell [2005], for the purpose of image matching. The significance of such special points lies in their compact representation of important image regions, leading to efficient indexing and good discriminative power, especially in object-based retrieval. In this domain, there has been a paradigm shift from global feature representations to local descriptors, as evidenced by a large number of recent publications. Typically, object categories or visual classes are represented by a combination of local descriptors and their spatial distributions, sometimes referred to collectively as part-based models. Variations usually arise out of the "prior" on the geometry imposed on the spatial relationship between the local parts, with extremes being fully independent (bag of features, each representing a part or region) and fully connected (constellation model [Fergus et al. 2003]). A fully connected model essentially limits the number of parts that can be modeled, since the algorithm complexity grows exponentially with it. As a compromise, sparser topologies have been proposed, such as the star topology [Fergus et al. 2005], a hierarchy with the lowest levels corresponding to local features [Bouchard and Triggs 2005], and a geometry where local features are spatially dependent on their nearest neighbors [Carneiro and Lowe 2006]. The model learning and categorization performance achieved in Fergus et al. [2003] has been improved upon, particularly in learning time, using contextual information and *boosting* in Amores et al. [2005, 2004]. A recent work [Zhang et al. 2006] uses segmentation to reduce the number of salient points for enhanced object representation. A discussion on the pros and cons of different types of color interest points used in image retrieval can be found in Gouet and Boujemaa [2002], while a comparative performance evaluation of the various proposed interest point detectors is reported in Mikolajczk and Schmid [2003]. The application of salient point detection for related feature extraction has also been explored. For example, interest point detectors have been employed for sparse texture representation, for the purpose of texture recognition, in Lazebnik et al. [2003].

Construction of Signatures from Features. In Figure 6, according to mathematical formulations, we summarize the types of signature roughly into vectors and distributions. As will be discussed in detail next, histograms and region-based signatures can both be regarded as sets of weighted vectors, and when the weights sum up to one, these sets are equivalent to discrete distributions (i.e., discrete in the sense that the support is finite). Our discussion will focus on region-based signature and its mathematical connection with histograms because it is the most exploited type of image signature. We note, however, that the distributions extracted from a collection of local feature vectors can be of other forms, for instance, a continuous density function [Do and Vetterli 2002], or even a spatial stochastic model [Li and Wang 2004]. A continuous density is generally more precise in describing a collection of local feature vectors than a discrete distribution with finitely many support vectors. A stochastic model moves beyond a continuous density by taking into account the spatial dependence among local feature vectors. For special kinds of images, we may need these sophisticated statistical models to characterize them. For instance, in Li and Wang [2004], it is noted that the spatial relationship among pixels is crucial for capturing Chinese ink painting styles. On the other hand, more sophisticated statistical models are computationally costly and less intuitive, a probable reason for their limited usage.

In earlier work, a histogram was a widely used form of distribution. Suppose the feature vectors are denoted by $x_{i,j} \in \mathcal{R}^d$, the d-dimensional Euclidean space. To form a basic histogram, $\mathcal{R}^d$ is divided into fixed bins and the percentage of $x_{i,j}$'s falling into each bin is calculated. Suppose there are k bins. A histogram can then be treated as a k-dimensional vector $(f_1, f_2, \ldots, f_k)^t$, where f_l is the frequency of the lth bin. Improvements over the basic histogram signature have been actively pursued. In Hadjidemetriou et al. [2004], a multiresolution histogram, together with its associated image matching algorithm, is shown effective in retrieving textured images. Computation of histograms at multiple resolutions maintains the simplicity and efficiency of ordinary histograms, but additionally captures spatial variations across images. In Jeong et al. [2004], Gaussian mixture vector quantization (GMVQ) is used to extract color histograms and shown to yield better retrieval than uniform- and vector quantization with squared error.

The disadvantages of treating histograms simply as vectors of frequencies are noted in Rubner et al. [1999]. The main issue is that the vector representation ignores the location of bins used to generate the histogram. For measuring the closeness of distributions, the locations of histogram bins are vital. The earth movers distance (EMD) is proposed in Rubner et al. [1999] to take into consideration bin locations. When EMD is used, the histogram is mathematically a collection of feature vector and frequency pairs: $\{(z_1, f_1), (z_2, f_2), \ldots, (z_k, f_k)\}$, where $z_l \in \mathcal{R}^d$ is the center or location of the lth bin. It is shown in Levina and Bickel [2001] that EMD, when applied to probability frequencies, is equivalent to the Mallows distance proposed in the early 1970's [Mallows 1972], which is a true metric for general probability measures. A histogram is a special distribution in the sense that it is discrete, that is, takes only countably many different values (for practical interest, finitely many). Moreover, histograms for different images are usually derived using a fixed set of bins.

Once the histogram is viewed as $\{(z_1, f_1), (z_2, f_2), \ldots, (z_k, f_k)\}$, namely a weighted set of vectors, a natural question to raise is why we have to employ a fixed set of bins located at $z_1, \ldots, z_k$. A direct extension from the histogram is to adpatively generate z_l and f_l together and also let the number of bins k depend on the image being handled. This is essentially the widely used region-based signature, as employed in Deng et al. [2001] and Wang et al. [2001]. Consider the dataset $\{x_{i,j}, 1 \le i, 1 \le j\}$. Applying a clustering procedure, for example, k-means, to the dataset groups the feature vectors $x_{i,j}$ into $\tilde{k}$ clusters such that feature vectors in the same cluster tend to be tightly packed. Let the mean of $x_{i,j}$'s in the same cluster l be z'_l. We thus have acquired a summary of the dataset: $\{(z'_1, f'_1), \ldots, (z'_{k'}, f'_{k'})\}$, where f'_l is the percentage of $x_{i,j}$'s grouped into cluster l. The collection of pixels (i, j) for which $x_{i,j}$'s are in the same cluster forms a relatively homogeneous region because the common cluster forces closeness between the visual features in $x_{i,j}$'s. This is why clustering of local feature vectors is a widely used method to segment images, and also why we call the signature $\{(z'_1, f'_1), \ldots, (z'_{k'}, f'_{k'})\}$ region-based.

With fixed bins, histograms of image feature vectors tend to be sparse in multidimensional space. In comparison, the region-based signature provides more compact description of images because it allows the representative vectors z'_l to adapt to images. In Deng et al. [2001] and Wang et al. [2001], it is argued that a region-based signature is more efficient computationally for retrieval, and that it also gets around the drawbacks associated with earlier proposals such as dimension reduction and color moment descriptors. Strictly speaking, a region-based signature is not merely a dynamic histogram representation, and, despite the mathematical connections made before, is not necessarily motivated by the intention of generalizing histograms. The motivation for using a region-based signature, as argued in Wang et al. [2001], is that a relatively homogeneous region of color and texture is likely to correspond to an object in an image. Therefore, by extracting regions, we crudely obtain a collection of objects, and with the objects in an image listed, it is easier to engage intuitions for defining similarity measures. Moreover, although we have z'_l, namely the mean of $x_{i,j}$'s in region l as a natural result of clustering, the description of the region can be expanded to include features not contained in z'_l, for instance, shape, which can only be meaningfully computed after the region has been formed.

Adaptive Image Signature. It is quite intuitive that the same set of visual features may not work equally well to characterize, say, computer graphics and photographs. To address this issue, learning methods have been used to tune signatures either based on images alone or by learning on-the-fly from user feedback. In Figure 6, we categorize image signatures according to their adaptivity into static, image-wise adaptive, and user-wise adaptive. Static signatures are generated in a uniform manner for all the images.

Image-wise adaptive signatures vary according to the classification of images. The term semantic-sensitive, coined in Wang et al. [2001], reflects such a mechanism to adjust signatures, and is a major trait of the SIMPLIcity system in comparison to its predecessors. Specifically, images are classified into several types first, and then signatures are formed from different features for these types. Despite the appeal of semantic-sensitive retrieval as a general framework, the classification conducted in SIMPLIcity only involves a small number of preselected image types (graph versus photograph, textured versus nontextured). The classification method relies on prior knowledge rather than training, and hence is not set-up for extension. More recently, semantic-sensitive features have also been employed in a physics-motivated approach [Ng et al. 2005], where images are distinguished as either photo-realistic rendering or photograph.

Care must be taken to ensure that the added robustness provided by heterogeneous feature representation does not compromise on the efficiency of indexing and retrieval. When a large number of image features are available, one way to improve generalization and efficiency is to work with a feature subset, or to impose different weights on the features. To avoid a combinatorial search, an automatic feature subset selection algorithm for SVMs is proposed in Weston et al. [2000]. Some of the other recent, more generic feature selection proposals involve boosting [Tieu and Viola 2004], evolutionary searching [Kim et al. 2000], Bayes classification error [Carneiro and Vasconcelos 2005], and feature dependency/similarity measures [Mitra et al. 2002]. An alternative way of obtaining feature weights based on user logs has been explored in Muller et al. [2004]. A survey and performance comparison of some recent algorithms on the topic can be found in Guyon and Elisseeff [2003].

Discussion. The various methods for visual signature extraction come with their share of advantages and limitations. While global features give the "big picture," local features represent the details. Therefore, depending on the scale of the key content or pattern, an appropriate representation should be chosen. In this sense, hybrid representations may sometimes be more attractive, but this may come at the cost of additional complexity. While segmentation is intended to recognize objects in a scene, precise segmentation still remains an open problem. Therefore, alternative approaches to characterize structure may be more suitable. However, such a representation may lose the charm of clear interpretability. Among different approaches to segmentation, there is often a tradeoff between quality and complexity which might lead to a difference in eventual search performance and speed. Hence, the choice of image signature to be used should depend on the desirability of the system.

In contrast with early years (Section 1.1), we have witnessed a major shift from global feature representations for images, such as color histograms and global shape descriptors, to local features and descriptors, such as salient points, region-based features, spatial model features, and robust local shape characterizations. It is not hard to imagine this shift to have been triggered by a realization that the image domain is too deep for global features to reduce the semantic gap. Local features often correspond with more meaningful image components, such as rigid objects and entities, which make association of semantics with image portions straightforward. The future in image feature- or signature representation resides both in theory and practice. Many years of research has made it clear that emulating human vision is very challenging, but that nonetheless, practical approaches can help to build useful systems. While the endeavor to characterize vision will likely continue, particularly in the core field of computer vision, practical approaches (e.g., fusion of local and global representations for top-down as well as a bottom-up representations) will potentially improve retrieval performance and user satisfaction in such systems. The availability of 3D and stereo image data, whenever obtainable, should be exploited to extract features more coherent to the human vision system. In summary, reducing the sensorial gap in tandem with the semantic gap should continue be a goal for the future.

3.2. Image Similarity Using Visual Signature

Once a decision on the choice of image signatures is made, how to use them for accurate image retrieval is the next concern. There has been a large number of fundamentally different frameworks proposed in recent years. Some of the key motivating factors behind the design of the proposed image similarity measures can be summarized as follows:

—agreement with semantics;

—robustness to noise (invariant to perturbations);

—computational efficiency (ability to work in real time and in large scale);

—invariance to background (allowing region-based querying); and

—local linearity (i.e., following triangle inequality in a neighborhood).

The various techniques can be grouped according to design philosophy as follows:

—treating features as vectors, nonvector representations, or ensembles;

—using region-based similarity, global similarity, or a combination of both;

—computing similarities over linear space or nonlinear manifold;

—considering the role played by image segments in similarity computation;

—using stochastic, fuzzy, or deterministic similarity measures; and

—use of supervised, semi-supervised, or unsupervised learning.

Leaving out those methods discussed in Smeulders et al. [2000], here we focus on some of the more recent approaches to image similarity computation.

Figure 7 shows the basic types of signature, distances ("dissimilarity measures") exploited, and underlying techniques needed to calculate these distances. For each type of signature, we also elucidate on its mathematical representation, which to a large extent determines the choice of distances and the employment of related methodologies. We will start our discussion on the region-based signature, since its widespread use is occurring in the current decade. The technical emphasis on region-based signature

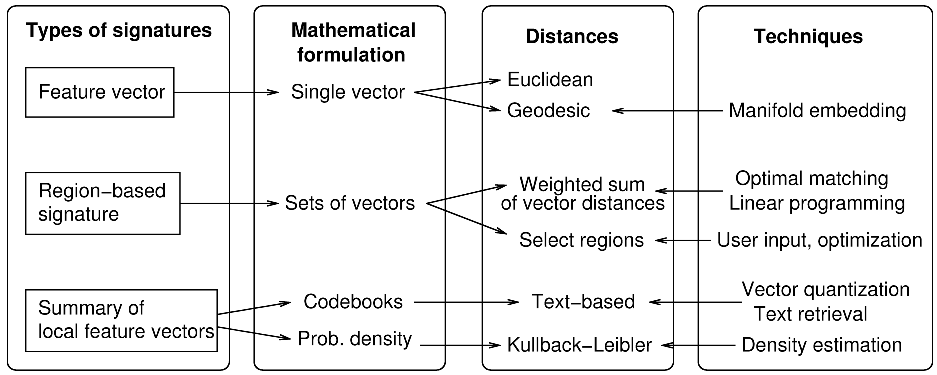


Fig. 7. Different types of image similarity measure, their mathematical formulations, and techniques for computing them.

rests on the definition of distance between sets of vectors, which is not as obvious as defining distance between single vectors. Research on this problem is further enriched by the effort to optimally choose a subset of regions pertaining to users' interests, thereby increasing robustness against inaccurate segmentation. Although global feature vectors had already been extensively used in the early years of CBIR, advances were achieved in recent years by introducing state-of-the-art learning techniques such as manifold embedding. Research efforts have been made to search for nonlinear manifolds in which the geodesic distances may better correspond to human perception. Instead of describing an image by a set of segmented regions, summaries of local feature vectors such as codebook and probability density functions have been used as signatures. Codebooks are generated by vector quantization, and the codewords are sometimes treated symbolically with application of text retrieval techniques. An effective way to obtain a density estimation is by fitting a Gaussian mixture model [Hastie et al. 2001], and the Kullback-Leibler distance is often used to measure the disparity between distributions.

First consider an image signature in the form of a weighted set of feature vectors $\{(z_1, p_1), (z_2, p_2), \ldots, (z_n, p_n)\}$, where z_i's are the feature vectors and p_i's are the corresponding weights assigned to them. The region-based signature discussed previously bears such a form, so a histogram can be represented in this way. Let us denote two signatures by $I_m = \{(z_1^{(m)}, p_1^{(m)}), (z_2^{(m)}, p_2^{(m)}), \ldots, (z_{n_m}^{(m)}, p_{n_m}^{(m)})\}$, $m = 1, 2$. A natural approach to defining a region-based similarity measure is to match $z_i^{(1)}$'s with $z_i^{(2)}$'s and then to combine the distances between these vectors as a distance between sets of vectors.

One approach to matching [Wang et al. 2001] is by assigning a weight to every pair $z_i^{(1)}$ and $z_j^{(2)}$, $1 \leq i \leq n_1$, $1 \leq j \leq n_2$, and the weight $s_{i,j}$ indicates the significance of associating $z_i^{(1)}$ with $z_j^{(2)}$. One motivation for soft matching is to reduce the effect on retrieval of inaccurate segmentation. The weights are subject to constraints, the most common ones being $\sum_i s_{i,j} = p_j^{(2)}$ and $\sum_j s_{i,j} = p_i^{(1)}$. Once the weights are determined, the distance between I_1 and I_2 is aggregated from the pair-wise distances between individual vectors. Specifically,

$$D(I_1, I_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} s_{i,j} d\left(z_i^{(1)}, z_j^{(2)}\right), \tag{1}$$

where the vector distance $d(\cdot,\cdot)$ can be defined in diverse ways depending on the system. Other matching methods include the Hausdorff distance, where every $z_i^{(1)}$ is matched to its closest vector in I_2, say $z_{i'}^{(2)}$, and the distance between I_1 and I_2 is the maximum among all $d(z_i^{(1)}, z_{i'}^{(2)})$. The Hausdorff distance is symmetricized by additionally computing the distance with the role of I_1 and I_2 reversed and choosing the larger of the two distances.

$$D_H(I_1, I_2) = \max\left(\max_i \min_j d\left(z_i^{(1)}, z_j^{(2)}\right),\ \max_j \min_i d\left(z_j^{(2)}, z_i^{(1)}\right)\right). \tag{2}$$

The Hausdorff distance is used for image retrieval in Ko and Byun [2002].

One heuristic to decide the matching weights $s_{i,j}$ for the pair $(z_i^{(1)}, z_j^{(2)})$ is to seek $s_{i,j}$'s such that $D(I_1, I_2)$ in Eq. (1) is minimized, subject to certain constraints on $s_{i,j}$. Suppose $\sum_i p_i^{(1)} = 1$ and $\sum_j p_j^{(2)} = 1$. This can always be made true by normalization, as long as there is no attempt to assign one image an overall higher signficance than the other. In practice, $p_i^{(1)}$'s (or $p_j^{(2)}$'s) often correspond to probabilities and automatically yield unit sum. Since $p_i^{(1)}$ indicates the significance of region $z_i^{(1)}$ and $\sum_j s_{i,j}$ reflects the total influence of $z_i^{(1)}$ in the calculation of $D(I_1, I_2)$, it is natural to require $\sum_j s_{i,j} = p_i^{(1)}$, for all i, and similarly $\sum_i s_{i,j} = p_j^{(2)}$, for all j. Additionally, we have the basic requirement $s_{i,j} \geq 0$ for all i, j. The definition of the distance is thus

$$D(I_1, I_2) = \min_{s_{i,j}} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} s_{i,j}\, d\left(z_i^{(1)}, z_j^{(2)}\right), \tag{3}$$

subject to $\sum_j s_{i,j} = p_i^{(1)}$, for all i, $\sum_i s_{i,j} = p_j^{(2)}$, for all j, and $s_{i,j} \geq 0$ for all i, j. This distance is precisely the Mallows distance in the case of discrete distributions [Mallows 1972].

The earth mover's distance (EMD) [Rubner et al. 2000] proposed early in the decade represents another soft matching scheme for signatures in the form of sets of vectors. The measure treated the problem of image matching as one of moving components of the color histograms of images from one to the other, with minimum effort, synonymous with moving earth piles to fill holes. When p_i and p'_j are probabilities, EMD is equivalent to the Mallows distance. Another useful matching-based distance is the IRM (integrated region matching) distance [Li et al. 2000]. The IRM distance uses the most similar highest priority (MSHP) principle to match regions. The weights $s_{i,j}$ are subject to the same constraints as in the Mallows distance, but $D(I_1, I_2)$ is not computed by minimization. Instead, the MSHP criterion entails that the pair of regions across two images with the smallest distance among all region pairs ought to be given the highest priority in matching, that is, to be assigned a maximum valid weight $s_{i,j}$. The matching is conducted recursively until all the region weights are consumed, namely until $\sum_j s_{i,j} = p_i^{(1)}$ and $\sum_i s_{i,j} = p_j^{(2)}$ have been achieved for all i and j. IRM is significantly faster to compute than the Mallows distance and has been found to not inferior, if not better than, the Mallows distance in terms of retrieval results.

Improvements over the basic matching idea have been made from different perspectives. These include tuning features according to image type, choosing region weights in more sophisticated ways, improving robustness against inaccurate segmentation, and speeding-up retrieval. In the SIMPLIcity system [Wang et al. 2001], a preliminary categorization (e.g., graph versus photograph, textured versus nontextured) is applied to images and different sets of features are used for each category. Region-based image retrieval, under the assumption of a hidden semantic concept underlying image generation, is explored in Zhang and Zhang [2004]. Here, a uniform, sparse region-based *visual dictionary* is obtained using self-organizing map (SOM)-based quantization, and images/regions are assumed *generated* probabilistically, conditional on hidden or latent variables that reflect on their underlying semantics. A framework for region-based image retrieval, with particular focus on efficiency, is proposed in Jing et al. [2004a]. Here, vector quantization (VQ) is employed to build a region codebook from training images, each entry sparsely or compactly represented, with the distinct advantages of efficiency and effectiveness in each case. To further speed-up retrieval, a tree-structured clustering is applied to images to narrow down the search range [Du and Wang 2001]. The system first uses a relatively simple signature, specifically a vector, to decide which cluster an image belongs to, and then uses the region-based signature and the IRM distance to compare the query with images in the chosen cluster.

A variation of IRM is attempted in Chen and Wang [2002] that employs fuzziness to account for inaccurate segmentation, to a greater extent. A new representation for object retrieval in cluttered images, without relying on accurate segmentation, is proposed in Amores et al. [2004]. Here, image model learning and categorization are improved upon using contextual information and boosting algorithms. A windowed search over location and scale is shown more effective in object-based image retrieval than methods based on inaccurate segmentation [Hoiem et al. 2004]. A hybrid approach involves the use of rectangular blocks for coarse foreground/background segmentation on the user's query region-of-interest (ROI), followed by a database search using only the foreground regions [Dagli and Huang 2004].

Without user input, image similarity measures usually attempt to take all the regions in an image into consideration. This may not be the best practice when users' interest is more specifically indicated than an example query image. For instance, if the query is a sketch drawn by a user, it may be meaningless to let the excluded areas of the sketch affect image comparison. It can be more desirable to match the sketch only to a relevant subset of regions automatically determined by the retrieval system, as explored in Ko and Byun [2002].

Even if the user starts searching with an example query image, it is sometimes assumed that he or she is willing to specify a portion of the image to be of interest. This argument has led to the concept of region-based querying. The Blobworld system [Carson et al. 2002], instead of performing image-to-image matching, lets users select one or more homogeneous color-texture segments, or *blobs*, as region(s)-of-interest. For example, if one or more segmented blobs identified by the user roughly correspond to a typical "tiger," then her search becomes equivalent to searching for the "tiger" object within images. For this purpose, the pictures are segmented into blobs using the E-M algorithm, and each blob b_i is represented as a color-texture feature vector $\mathbf{v_i}$. Given a query blob b_i, and every blob b_j in the database, the most similar blob has score

$$\mu_i = \max_j \exp\left(\frac{(\mathbf{v_i} - \mathbf{v_j})^{\mathbf{T}} \Sigma (\mathbf{v_i} - \mathbf{v_j})}{2}\right), \tag{4}$$

where matrix Σ corresponds to user-adjustable weights on specific color and texture features. The similarity measure is further extended to handle compound queries using fuzzy logic. While this method can lead to more precise formulation of user queries and can help users to better understand the computer's responses, it also requires greater involvement from and dependence on them. For finding images containing scaled or translated versions of query objects, retrieval can also be performed without any explicit involvement of the user [Natsev et al. 2004].

As discussed previously, regions are obtained by segmenting images using local feature vectors. Roughly speaking, region-based signatures can be regarded as a result of summarizing these feature vectors. Along the lines of using a summary of local feature vectors as the signature, there are other approaches explored. For instance, in Iqbal and Aggarwal [2002], primitive image features are hierarchically and perceptually grouped and their interrelationships used to characterize structure [Iqbal and Aggarwal 2002]. Another approach is the use of vector quantization (VQ) on image blocks to generate codebooks for representation and retrieval, taking inspiration from data compression and text-based strategies [Zhu et al. 2000]. For textured images, segmentation is not critical. Instead, distributions of the feature vectors are estimated and used as signatures. Methods for texture retrieval using the Kullback-Leibler (K-L) divergence have been proposed in Do and Vetterli [2002] and Mathiassen et al. [2002]. The K-L divergence, also known as the *relative entropy*, is an asymmetric information-theoretic measure of the difference between two distributions $f(\cdot)$ and $g(\cdot)$, defined as

$$K(f,g) = \int_{-\infty}^{+\infty} f(x) log \frac{f(x)}{g(x)} dx, \qquad K(f,g) = \sum_x f(x) \log \frac{f(x)}{g(x)} \tag{5}$$

in the continuous and discrete cases, respectively. Fractal-block-code-based image histograms have been shown effective in retrieval on texture databases [Pi et al. 2005]. The use of the MPEG-7 content descriptors to train self-organizing maps (SOMs) for image retrieval is explored in Laaksonen et al. [2002].

When images are represented as single vectors, many authors note an apparent difficulty in measuring perceptual image distance by metrics in any given *linear* feature space. One approach to tackle this issue is to search for a nonlinear manifold in which the image vectors lie, and to replace the Euclidean distance by the geodesic distance. The assumption here is that visual perception corresponds better with this nonlinear subspace than with the original linear space. Computation of similarity may then be more appropriate if performed nonlinearly along the manifold. This idea is explored and applied to image similarity and ranking in He [2004], Vasconcelos and Lippman [2005], He et al. [2004a, 2004b, 2004c], and Zhou et al. [2003]. Typical methods for learning underlying manifolds, which essentially amounts to nonlinear dimension reduction, are locally-linear embedding (LLE), isomapping, and multidimensional scaling (MDS) [de Silva and Tenenbaum 2003].

The different distance measures discussed so far have their own advantages and disadvantages. While simple methods lead to very efficient computation, which in turn makes image ranking scalables (a quality that greatly benefits real-world applications), they often are not effective enough to be useful. Depending on the specific application and image signatures constructed, a very important step in the design of an image retrieval system is the choice of distance measure. Factors that differ across various distance measures include type of input, method of computation, computational complexity, and whether the measure is a metric. In Table I, we summarize distance measures according to these factors, for ease of comparison.

In the previous subsection, we discussed tuning image signatures by categorizing images or by learning from user preferences. A closely related issue is to tune image similarity measures. It is, in fact, impossible to completely set apart the two types of adaptivity, since tuning signatures ultimately results in a change of similarity. Referring a tuning method in one way or another is often merely a matter of whichever is easier to understand. Automatic learning of image similarity measures with the help of contextual information has been explored in Wu et al. [2005]. In the case that a valid pair-wise image similarity metric exists despite the absence of an explicit vectored

Table I. Popular Distance Measures Used for Similarity Computation in Image Retrieval

Distance Measure	Input	Computation	Complexity	Metric	Comments
Euclidean (L^2norm)	$\vec{X}_a, \vec{X}_b \in \mathrm{R}^n$ (vectors)	$\vec{X}_a \cdot \vec{X}_b$	$\Theta(n)$	Yes	Popular, fast, L^1 also used
Weighted Euclidean	$\vec{X}_a, \vec{X}_b \in \mathrm{R}^n$ $W \in \mathrm{R}^n$ (vec. + wts.)	$\vec{X}_a^T[W]\vec{X}_b$ $[\cdot] \leftarrow$ diagonalize	$\Theta(n)$	Yes	Allows features to be weighted
Hausdorff	Vector sets: $\{\vec{X}_a^{(1)}, .., \vec{X}_a^{(p)}\}$ $\{\vec{X}_b^(1), .., \vec{X}_b^{(q)}\}$	See Eqn. 2	$\Theta(pqn)$ $(d(\cdot,\cdot) \leftarrow$ L^2 norm)	Yes	Sets corr. to image segments
Mallows	Vector sets: $\{\vec{X}_a^{(1)}, .., \vec{X}_a^{(p)}\}$ $\{\vec{X}_b^(1), .., \vec{X}_b^{(q)}\}$ Signific.: S	See Eqn. 3	$\Theta(pqn)$ + variable part	Yes	The EMD is its special case
IRM	Vector sets: $\{\vec{X}_a^{(1)}, .., \vec{X}_a^{(p)}\}$ $\{\vec{X}_b^(1), .., \vec{X}_b^{(q)}\}$ Signific.: S	See Eqn. 3	$\Theta(pqn)$ + variable part	No	Much faster than Mallows computation in practise
K-L divergence	$\vec{F}, \vec{G} \in \mathrm{R}^m$ (histograms)	$\sum_x F(x) \log \frac{F(x)}{G(x)}$	$\Theta(m)$	No	Asymmetric, compares distributions

representation in some metric space, *anchoring* can be used for ranking images [Natsev and Smith 2002]. Anchoring involves choosing a set of representative *vantage* images, and using the similarity measure to map an image into a vector. Suppose there exists a valid metric $d(F_i, F_j)$ between each image pair, and a chosen set of K vantage images $\{A_1, \ldots, A_K\}$. A *vantage space transformation* $V : F \rightarrow \mathcal{R}^K$ then maps each image F_i in the database to a vectored representation $V(F_i)$ as follows.

$$V(F_i) =< d(F_i, A_1), \ldots, d(F_i, A_K) > \tag{6}$$

With the resultant vector embedding, and after similarly mapping a query image in the same space, standard ranking methods may be applied for retrieval. When images are represented as ensembles of feature vectors, or underlying distributions of the low-level features, visual similarity can be ascertained by means of nonparametric tests such as Wald-Wolfowitz [Theoharatos et al. 2005] and K-L divergence [Do and Vetterli 2002]. When images are conceived as bags of feature vectors corresponding to regions, multiple-instance learning (MIL) can be used for similarity computation [Zhang et al. 2002].

A number of probabilistic frameworks for CBIR have been proposed in the last few years [Jin and Hauptmann 2002; Vasconcelos and Lippman 2000b]. The idea in Vasconcelos and Lippman [2000b] is to integrate feature selection, feature representation, and similarity measures into a combined Bayesian formulation, with the objective of minimizing the probability of retrieval error. One problem with this approach is the computational complexity involved in estimating probabilistic similarity measures. The complexity is reduced in Vasconcelos [2004] using VQ to approximately model the probability distribution of the image features.

Discussion. As shown in Figure 7, similarity computation can be performed with feature vectors, region-based signatures, or summarized local features. The main advantage of a single vector representing an image is that algebraic and geometric operations can be performed efficiently and in a principled fashion. However, many such representations lack the necessary detail to represent complex image semantics. For example, a picture of two cups on a plate by a windowsill cannot easily be mapped to a finite vector representation, simply because the space of component semantics is extremely large, in practice. Instead, if a concatenation of region descriptors is used to represent a picture, it is more feasible to map the component semantics (e.g., cup, window) to image regions. On the other hand, extracting semantically coherent regions is in itself very challenging. Probabilistic representations can potentially provide an alternative, allowing rich descriptions with limited parametrization.

The early years of research (Section 1.1) showed us the benefits as well as the limitations of feature vector representations. They also paved the way for the new breed of region-based methods, which have now become more standard than ever before. The idea of region-based image querying also gained prominence in the last few years. Many new salient-feature-based spatial models were introduced, particularly for recognizing objects within images, building mostly on early, pre-2000 work. The idea that image similarity is better characterized by geodesic distances over a nonlinear manifold embedded in the feature space has improved upon earlier notions of a linear embedding of images. A number of systems have also been introduced for public usage in recent years. The future of image similarity measures lies in many different avenues. The subjectivity in similarity needs to be incorporated more rigorously into image similarity measures, to achieve what can be called *personalized* image search. This can also potentially incorporate ideas beyond the semantics, such as aesthetics and personal preferences in style and content. Extensions of the idea of nonlinear image manifolds to incorporate the whole spectrum of natural images, and to allow adaptability for personalization, are avenues to consider. While development of useful systems continues to remain critical, the ever-elusive problem of reducing the semantic gap needs concerted attention.

3.3. Clustering and Classification

Over the years it has been observed that it is too ambitious to expect a single similarity measure to produce robust, perceptually meaningful ranking of images. As an alternative, attempts have been made to augment the effort with learning-based techniques. In Table II, for both clustering and classification, we summarize the augmentations to traditional image-similarity-based retrieval, the specific techniques exploited, and the limitations, respectively.

Image classification or categorization has often been treated as a preprocessing step for speeding-up image retrieval in large databases and improving accuracy, or for performing automatic image annotation. Similarly, in the absence of labeled data, unsupervised clustering has often been found useful for retrieval speedup as well as improved result visualization. While image clustering inherently depends on a similarity measure, image categorization has been performed by varied methods that neither require nor make use of similarity metrics. Image categorization is often followed by a step of similarity measurement, restricted to those images in a large database that belong to the same visual class as predicted for the query. In such cases, the retrieval process is intertwined, whereby categorization and similarity matching steps together form the retrieval process. Similar arguments hold for clustering as well, due to which, in many cases, it is also a fundamental “early” step in image retrieval.

In recent years, a considerable number of innovations have been accomplished for both clustering and classification, with tremendously diverse target applications. It is not our intention here to provide a general review of these technologies. We refer to Hastie et al. [2001] for the basic principles and a more comprehensive review. We will restrict ourselves to new methods and applications appearing in image retrieval and closely related topics.

Table II. Comparison of Three Different Learning Techniques in Their Application to Image Retrieval

Augmentation (User Involvement)	Purpose	Techniques	Drawbacks
Clustering (minimal)	Meaningful result visualization, faster retrieval, efficient storage	Side-information, kernel mapping, k-means, hierarchical, metric learning [Chen and Wang 2004] [Hastie et al. 2001] [Sebe et al. 2000] [Wu et al. 2005]	Same low-level features, poor user adaptability
Classification (requires prior training data, not interactive)	Pre-processing, fast/ accurate retrieval, automatic organization	SVM, MIL, statistical models, Bayesian classifiers, k-NN, trees [Zhang et al. 2002] [Hastie et al. 2001] [Panda and Chang 2006]	Training introduces bias, many classes unseen
Relevance Feedback (significant, interactive)	Capture user and query specific semantics, refine rank accordingly	Feature re-weighting, region weighting, active learning, memory/ mental retrieval, boosting [Hastie et al. 2001] [Rui et al. 1998] [Jaimes et al. 2004] [Fang and Geman 2005]	Same low level features, increased user involvement

Unsupervised clustering techniques are a natural fit when handling large, unstructured image repositories such as the Web. Figure 8 summarizes clustering techniques according to the principles of clustering and shows the applicability of different methods when the mathematical representation of learning instances varies. Again, we divide the instances into three types: vectors, sets of vectors, and stochastic processes (including distributions); these are consistent with the categorization of image signatures discussed in the previous subsection. From the perspective of the application, clustering specifically for Web images has received particular attention from the multimedia community, where metadata is often available for exploitation, in addition to visual features [Wang et al. 2004a; Gao et al. 2005; Cai et al. 2004].

Clustering methods fall roughly into three types: pair-wise-distance-based, optimization of an overall clustering quality measure, and statistical modeling. The pair-wise-distance-based methods (e.g., linkage clustering and spectral graph partitioning) are of general applicability, since the mathematical representation of the instances becomes irrelevant. They are particularly appealing in image retrieval because image signatures often have complex formulation. One disadvantage, however, is the high computational cost because we need to compute an order of n^2 pair-wise distances, where n is the size of the dataset. In Zheng et al. [2004], a locality-preserving spectral clustering technique is employed for image clustering in such a way that unseen images can be placed into clusters more easily than with traditional methods. In CBIR systems, which retrieve images ranked by relevance to the query image only, similarity information among the retrieved images is not considered. In this respect, Chen et al. [2005] proposes the use of a new spectral clustering- [Shi and Malik 2000] based approach to incorporate such information into the retrieval process. In particular, clusters are dynamically generated, tailored specifically to the query image each time, to improve retrieval performance.

Clustering based on the optimization of an overall measure of clustering quality is a fundamental approach explored since the early days of pattern recognition work. The immensely popular, k-means clustering method is one example. In k-means, the merit of a clustering result is measured by the sum of within-cluster distances between every vector and its cluster centroid. This criterion ensures that the clusters generated

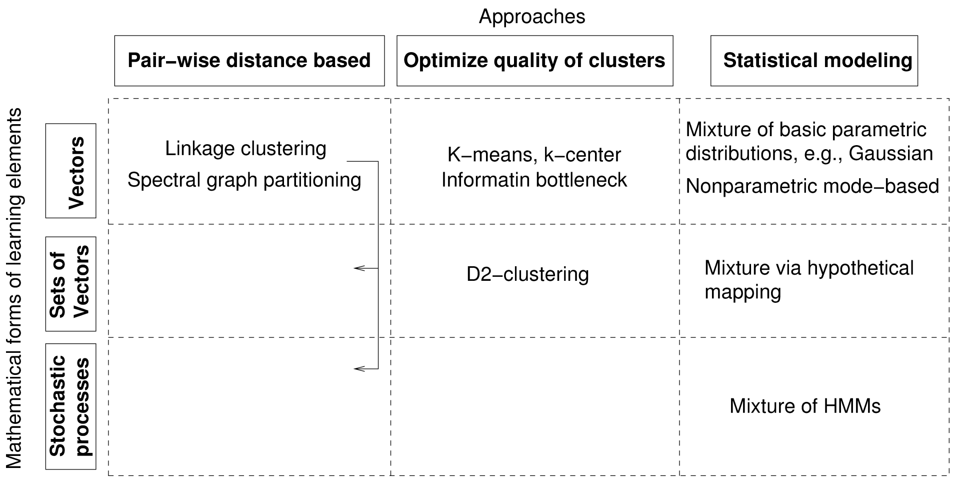


Fig. 8. Paradigms of clustering methods and their scope of applications.

are tight, a generally accepted heuristic. Here, if the number of clusters is not specified, a simple method to determine this number is to gradually increase it until the average distance between a vector and its cluster centroid is below a given threshold. A more sophisticated way to determine the number of clusters is the competitive agglomeration algorithm, with application to image clustering [Saux and Boujemaa 2002]. In Gordon et al. [2003], an unsupervised clustering approach for images has been proposed using the information bottleneck (IB) principle. The proposed method works for discrete (histograms) as well as for continuous (Gaussian mixture) image representations. Clustering based on the IB principle [Tishby et al. 1999] can be summarized as follows: Given two variables A (which we try to compress/cluster) and B (which contains relevant information), and their joint distribution $Pr(A, B)$, we seek to perform soft partitioning of A by a probabilistic mapping V, namely, $Pr(V|A)$, such that the mutual information among A and V is minimized, while the relevant information among B and V is maximized.

In k-means clustering, a centroid vector is computed for every cluster. This centroid vector is chosen to minimize the sum of within-cluster distances. When the Euclidean distance is used, it can easily be shown that the centroid ought to be the average of the vectors in a cluster. For nonvector data, the determination of the centroid can be challenging. The extension of k-means to instances represented by sets of weighted vectors is made in Li and Wang [2008] by means of the D2-clustering algorithm. The Mallows distance is used for region-based image signatures represented as sets of weighted arbitrary vectors. When the weights assigned to the vectors are probabilities, this representation is essentially a discrete distribution. The centroid for every cluster is also a discrete distribution, for which both the probabilities and vectors in the support domain need to be solved. Although D2-clustering uses the same intrinsic clustering criterion of as k-means, computationally, it is much more complex due to the complexity of the instances themselves. Large-scale linear programming is used for the optimization in D2-clustering. Another algorithm for clustering sets of vectors is developed using the IRM distance [Li 2005]. As compared with D2-clustering, this algorithm is similar in principle and significantly faster, but has weaker optimization properties.

Statistical modeling is another important paradigm for clustering. The general idea is to treat every cluster as a pattern characterized by a relatively restrictive distribution, and the overall dataset is thus a mixture of these distributions. For continuous vector data, the most commonly used distribution of individual vectors is the Gaussian distribution. By fitting a mixture of Gaussians to a dataset, usually by the EM algorithm [McLachlan and Peel 2000], we estimate the means and covariance matrices of the Gaussian components, which correspond to the center locations and shapes of clusters. One advantage of the mixture modeling approach is that it not only provides a partition of data, but also yields an estimated density, which sometimes is itself desired [Do and Vetterli 2002]. The component in a mixture model is not always a multivariate distribution. For instance, in Li and Wang [2004], the objects to be clustered are large areas of images, and every cluster is characterized by a 2D MHMM. As long as a probability measure can be set-up to describe a cluster, the mixture modeling approach applies seamlessly. When it is difficult to form a probability measure in a certain space, a mixture model can be established by clustering the data and mapping each cluster to a distance-preserving Euclidean space [Li and Wang 2008]. In this case, the mixture model is not used to yield clustering, but to better represent a dataset, eventually resulting in better classification.

Image categorization (classification) is advantageous when the image database is well specified, and labeled training samples are available. Domain-specific collections such as medical image databases, remotely sensed imagery, and art and cultural image databases are examples where categorization can be beneficial. Classification is typically applied for either automatic annotation, or for organizing unseen images into broad categories for the purpose of retrieval. Here we discuss the latter. Classification methods can be divided into two major branches: discriminative and generative modeling approaches. In discriminative modeling, classification boundaries or posterior probabilities of classes are estimated directly, for example, SVM and decision trees. In generative modeling, the density of data within each class is estimated and the Bayes formula is then used to compute the posterior. Discriminative modeling approaches are more direct when optimizing classification boundaries. On the other hand, generative modeling approaches are easier to incorporate with prior knowledge and can be used more conveniently when there are many classes.

Bayesian classification is used for the purpose of image retrieval in Vailaya et al. [2001]. A textured/nontextured and graph/photograph classification is applied as preprocessing to image retrieval in Wang et al. [2001]. Supervised classification based on SVMs has been applied to images in Goh et al. [2001]. A more recent work describes an efficient method for processing multimedia queries in an SVM-based supervised learning framework [Panda and Chang 2006]. SVMs have also been used in an MIL framework in Chen and Wang [2004]. In the MIL framework, a set of, say, l training images for learning an image category are conceived as labeled bags $\{(B_1, y_1), \ldots, (B_l, y_l)\}$, where each bag B_i is a collection of instances $v_{ij} \in \mathbf{R}^m$. Each instance v_{ij} corresponds to a segmented region j of a training image i, and $y_i \in \{-1, +1\}$ indicating a negative or positive example with respect to the category in question. The key idea is to map these bags into a new feature space where SVMs can be trained for eventual classification. Image classification based on a generative model for the purpose of retrieval is explored in Datta et al. [2007].

Discussion. Clustering is a hard problem with two unknowns, namely, the number of clusters, and the clusters themselves. In image retrieval, clustering helps in visualization and retrieval efficiency. The usual problems of clustering-based applications appear here as well, whereby the clusters may not be representative enough or accurate for visualization. While supervised classification is more systematic, the availability of comprehensive training data is often scarce. In particular, the veracity of "ground truth" in image data itself is a subjective question.

Clustering and classification for the purpose of image retrieval received relatively less attention in the early years of research. The spotlight was on feature extraction and similarity computation. With the need for practical systems that scale well to billions of images and millions of users, practical hacks such as preclustering and fast classification have become critical. The popularization of new information-theoretic clustering methods, as well as classification methods such as SVM and boosting, have led to their extensive use in the image retrieval domain as well. New generative models such as latent Dirichlet allocation (LDA) and 2D-MHMM have made their way into image modeling and annotation. The future, in our opinion, lies in supervised and unsupervised generative models for characterizing the various facets of images and metadata. There is often a lot of structured and unstructured data available with the images that can be potentially exploited through joint modeling, clustering, and classification. It is difficult to guess how much these methods can help bridge the semantic or sensorial gap, but one fact is unequivocal: System implementations can greatly benefit in various ways from the efficiency that these learning-based methods can produce.

3.4. Relevance Feedback-Based Search Paradigms

The approach to search has an undeniable tie to the underlying core technology because it defines the goals and the means to achieve them. One way to look at the types of search is through modality (i.e., query by keyword/keyphrase, by example images, or a combination of both, as discussed in Section 2). Other ways to characterize search is by the nature and level of human and system interaction involved, and the user intent (Section 2). In this section, we concentrate on the latter categorization, exploring the different search paradigms that affect how humans interact and systems interpret/respond.

Relevance feedback (RF) is a query modification technique which attempts to capture the user's precise needs through iterative feedback and query refinement. It can be thought of as an alternative search paradigm, complementing other paradigms such as keyword-based search. Ever since its inception in the CBIR community [Rui et al. 1998], a great deal of interest in query modification has been generated. In the absence of a reliable framework for modeling high-level image semantics and subjectivity of perception, the user's feedback provides a way to learn case-specific query semantics. While a comprehensive review can be found in Zhou and Huang [2003], here we present a short overview of recent work in RF, and the various ways in which these advances can be categorized. We group them here based on the nature of the advancements made, resulting in (possibly overlapping) sets of techniques that have pushed the frontiers in a common domain. These include: (a) learning-based, (b) feedback specification, (c) user-driven, (d) probabilistic, (e) region-based, and (f) other advancements.

Learning-Based Advancements. Based on the user's relevant feedback, learning-based approaches are typically used to appropriately modify the feature set or similarity measure. However, in practice, a user's RF results in only a small number of labeled images pertaining to each high-level concept. This obstacle, along with other unique challenges pertiaining to RF, has generated interest in novel machine-learning techniques to solve the problem, such as *one-class* learning, *active* learning, and *manifold* learning. To circumvent the problem of learning from small training sets, a discriminant-EM algorithm is proposed to make use of unlabeled images in the database for selecting more discriminating features [Wu et al. 2000]. On the other hand, it is often the case that the positive examples received due to feedback are more consistently located in the feature space than are the negative examples, the latter of which may consist of any irrelevant image. This leads to a natural formulation of *one-class* SVM for learning relevant regions in the feature space from feedback [Chen et al. 2002]. Let $\{\mathbf{v_1}, \ldots, \mathbf{v_n}\}$, $\mathbf{v_i} \in \mathbf{R}^d$ be a set of n positive training samples. The idea is to find a mapping $\Phi(\mathbf{v}_i)$ such that most samples are tightly contained in a hypersphere of radius R in the mapped space, subject to regularization. The primal form of the objective function is thus given by

$$\min_{R,e,c} \left(R^2 + \frac{1}{kn} \sum_i e_i \right) \text{ subject to } \|\Phi(\mathbf{v_i}) - c\|^2 \leq R^2 + e_i, e_i \geq 0, i \in \{1, \ldots, n\}. \quad (7)$$

Here, c is the hypersphere center in the mapped space, and $k \in [0, 1]$ is a constant that controls the tradeoff between radius of the sphere and number of samples it can hold. Among other techniques, a principled approach to optimal learning from RF is explored in Rui and Huang [2000]. We can also view RF as an active learning process, where the learner chooses an appropriate subset for feedback from the user in each round based on her previous rounds of feedback, instead of choosing a random subset. Active learning using SVMs was introduced into RF in Tong and Chang [2001]. Extensions to active learning have also been proposed [Goh et al. 2004; He et al. 2004b]. In He et al. [2004b], it is conceived that image features reside on a manifold embedded in the Euclidean feature space. Under this assumption, relevant images to the query provided by RF, along with their nearest neighbors, are used to construct a subgraph over the images. The geodesic distances, that is, the shortest path on the graph between pairs of vertices representing image pairs, are then used to rank images for retrieval.

Feedback Specification Advancements. Traditionally, RF has engaged the user in multiple rounds of feedback, each round consisting of one set each of positive and negative examples in relation to the intended query. However, recent work has introduced other paradigms of query specification that have been found either more intuitive, or more effective. Feedback based directly on an image semantics characterized by manually defined image labels, appropriately termed *semantic feedback*, is proposed in Yang et al. [2005b]. A well-known issue with feedback solicitation is that multiple rounds of feedback test the user's patience. To circumvent this problem, user logs of earlier feedback can be used in query refinement, thus reducing the user engagement in RF, as shown in Hoi and Lyu [2004b]. Innovation has also come in the form of the nature by which feedback is specified by the user. In Kim and Chung [2003], the notion of a multipoint query, where multiple-image examples may be used as query and in the intermediate RF step, is introduced. At each round of the RF, clusters of images found relevant based on the previous feedback step are computed, whose representatives form the input for the next round of RF. It is well known that there is generally an asymmetry between the sets of positive and negative image examples presented by the user. In order to address this asymmetry during RF when treating it as a two-class problem, a biased discriminant-analysis-based approach has been proposed in Zhou and Huang [2001b]. While most algorithms treat RF as a two-class problem, it is often intuitive to consider multiple groups of images as relevant or irrelevant [Hoi and Lyu 2004a; Nakazato et al. 2003; Zhou and Huang 2001a]. For example, a user looking for "cars" can highlight groups of blue and red cars as relevant, since it may not be possible to represent the concept of "car" uniformly in a visual feature space. Another novelty in feedback specification is the use of multilevel relevance scores to indicate varying degrees of relevance [Wu et al. 2004].

User-Driven Methods. While many past attempts at RF have focused on the machine's ability to learn from user feedback, the user's point-of-view in providing the feedback has largely been taken for granted. Of late, there has been some interest in design RF paradigms aimed to help users. In some new developments, there have been attempts at tailoring the search experience by providing the user with cues and hints for more specific query formulation [Jaimes et al. 2004; Nagamine et al. 2004]. While the approach may still involve RF from the system point-of-view, it is argued that the human memory can benefit from provided cues for better query formulation. A similar search paradigm proposed in Fang and Geman [2005] and Fang et al. [2005] models successive user responses using a Bayesian, information-theoretic framework. The goal is to "learn" a distribution over the image database representing the mental image of the user, and to use this distribution for retrieval. Another well-known issue with a human in the loop is that multiple rounds of feedback are often bothersome for the user. This redundancy has been alleviated in Hoi and Lyu [2004b] by making use of logs that contain earlier feedback given by that user. Recently, a manifold learning technique to capture user preference over a *semantic manifold* from RF has been proposed in Lin et al. [2005].

Probabilistic Methods. Probabilistic models, while popular in the early years of image retrieval for tackling the basic problem, have found increasing patronage for performing RF in recent years. Probabilistic approaches have been taken in Cox et al. [2000], Su et al. [2003], and Vasconcelos and Lippman [2000a]. In Cox et al. [2000], the PicHunter system is proposed, where uncertainty about the user's goal is represented by a distribution over the potential goals, following which the Bayes' rule helps in selecting the target image. In Su et al. [2003], RF is incorporated using a Bayesian-classifier-based reranking of the images after each feedback step. The main assumption used here is that the features of positive examples, which potentially reside in the same semantic class, are all generated by an underlying Gaussian density. The RF approach in Vasconcelos and Lippman [2000a] is based on the intuition that the system's belief at a particular time about the user's intent is a *prior*, while the subsequent user feedback is *new* information obtained. Together, these concepts help to compute the new belief about the intent, using the Bayes' rule, which in turn becomes the prior for the next feedback round.

Region-Based Methods. With an increased popularity of region-based image retrieval [Carson et al. 2002; Wang et al. 2001; Ko and Byun 2002], attempts have been made to incorporate the *region* factor into RF. In Jing et al. [2004a], two different RF scenarios are considered, and retrieval is tailored to support each of them through query point modification and SVM-based classification, respectively. In this feedback process, the region importance (RI) for each segmented region is learned, for successively better retrieval. This core idea, namely that of integrating region-based retrieval with relevance feedback, has been further detailed for the two RF scenarios in Jing et al. [2004b].

Other Advancements. Besides the grouped sets of methods, there have been a number of isolated advancements covering various aspects of RF. For example, methods for performing RF, using visual as well as textual features (metadata) in unified frameworks, have been reported in Lu et al. [2000], Zhou and Huang [2002], Amores et al. [2004], and Jing et al. [2005]. A tree-structured SOM has been used as an underlying technique for RF [Laaksonen et al. 2001] in a CBIR system [Laaksonen et al. 2002]. A well-known RF problem regarding query specification is the fact that after each round of user interaction, the top query results need to be recomputed following some modification. A way to speed-up this *nearest-neighbor* search is proposed in Wu and Manjunath [2001]. The use of RF for helping to capture the relationship between low-level features and high-level semantics, a fundamental problem in image retrieval, has been attempted using logs of user feedbacks in Han et al. [2005].

Discussion. Relevance feedback provides a compromise between a fully automated, unsupervised system and one based on subjective user needs. While query refinement is an attractive proposal when considering a very diverse user base, there is also the question of how well the feedback can be utilized for refinement. Whereas a user would prefer shorter feedback sessions, there is an issue as to how much feedback is enough for the system to learn the user needs. One issue which has been largely ignored in past RF research is that the user's needs might evolve over the feedback steps, making weaker the assumption of a fixed target. New approaches such as Jaimes et al. [2004] and Fang and Geman [2005] have started incorporating this aspect of the user's mind in the RF process.

Relevance feedback was introduced into image retrieval at the end of the previous decade (Section 1.1). Today, it is a more mature field, spanning many different subtopics and addressing a number of practical concerns while keeping in mind the user in the loop. While this progress is evident, the issue remains that, we do not see many real-world implementations of the relevance feedback technology, either in the image or text retrieval domains. This is potentially due to the feedback process that the users must go through, which tests the users' patience. New ideas such as memory retrieval, which actually provide the user with benefits in the feedback process, may possibly be key to popularizing RF. The future of this field clearly lies in its practical applicability, focusing on how the user can be spared the greatest amount of effort in conveying the desired semantics. The breaking-point and utility derived out of this process, at which the user runs out of patience and at which she is satisfied with the response, respectively, must be studied for better system design.

3.5. Multimodal Fusion and Retrieval

Media relevant to the broad area of multimedia retrieval and annotation includes, but is not limited to, images, text, free text (unstructured, e.g., paragraphs), graphics, video, and any conceivable combination of them. Thus far, we have encountered a multitude of techniques for modeling and retrieving images, and text associated with these images. While not covered here, the reader may be aware of equally broad spectrums of techniques for text, video, music, and speech retrieval. In many cases, these independent, media-specific methods do not suffice to satiate the needs of users who are seeking what they can best describe only by a combination of media. Therein lies the need for multimodal fusion as a technique for satisfying such user queries. We consider this as one of the "core" techniques because in principal, it is distinct from any of the methods we have discussed so far. Even with the very good retrieval algorithms available independently for two different media, effectively combining them for multimodal retrieval may be far from trivial. Research in fusion learning for multimodal queries therefore attempts to learn optimal combination strategies and models.

Fortunately (for researchers) or unfortunately (for users), precious little multimodal fusion has been attempted in the context of image retrieval and annotation. This opens avenues for exploring novel user interfaces, querying models, and resulting visualization techniques pertinent to image retrieval, in combination with other media. Having said that, we must point out that multimodal fusion has indeed been attempted in more obvious problem settings within *video retrieval*. With this field as an example, we briefly expose readers to multimodal fusion, in the hope that it motivates image retrieval research that takes advantage of these techniques. We believe that the need for mutimodal retrieval in relation to images will soon grow in stature.

When video data comes with closed-captions and/or associated audio track, these components can prove to be useful items of metadata for retrieval as well. One of the key problems faced in video retrieval research is therefore the combination or fusion of responses from these multiple modalities. It has been observed and reported that multimodal fusion almost always enhances retrieval performance for video [Hauptmann and Christel 2004]. Usually, fusion involves learning some kind of combination rules across multiple decision streams (ranked lists or classifier responses) using a certain amount of data with ground truth as a validation set. This is also referred to as late fusion. Alternative approaches to fusion involve classifier retraining. In Wu et al. [2004], multimodal fusion has been treated as a two-step problem. The first step involves finding statistically independent modalities, followed by superkernel fusion to determine their optimal combination. Fusion approaches have been found beneficial for important video applications such as detection of documentary scene changes [Velivelli et al. 2004] and story segmentation [Zhai et al. 2005]. Fusion learning has been found to outperform naive fusion approaches, as well as the oracle (best performer) for the TRECVID 2005 query retrieval task [Joshi et al. 2007].

Discussion. Fusion learning is an offline process while fusion application at real time is computationally inexpensive. Hence multimodal fusion is an excellent method to boost retrieval performance at real time. However, special care needs to be taken to ensure that the fusion rules do not overfit the validation set used for learning them. Usually, data resampling techniques such as bagging are found to help avoid overfitting, to some extent. Fusion techniques can also be used to leverage classifiers built for numerous concepts with possible semantic coherence, whether the underlying data is image or video.

Fusion for image retrieval is a fairly novel area, with very little achieved in the early days of research. The ideas of fusion go hand-in-hand with practical, viable system development, which is critical for the future of image retrieval research. We live in a truly multimedia world, and we as humans always take the benefit of each media for sensory interpretation (see, hear, smell, taste, touch). There is no reason why advantage should not be taken of all available media (images, video, audio, text) for building useful systems. The future lies in harnessing as many channels of information as possible, and fusing them in smart, practical ways to solve real problems. Principled approaches to fusion, particularly probabilistic ones, can also help provide performance guarantees which in turn convert to quality standards for public-domain systems.

4. CBIR OFFSHOOTS: PROBLEMS AND APPLICATIONS OF THE NEW AGE

Smeulders and coauthors [Smeulders et al. 2000] surveyed CBIR at the end of what they referred to as the early years of its study. The field was presented as a natural successor to certain existing disciplines such as computer vision, information retrieval, and machine learning. However, in the last few years, CBIR has evolved and emerged as a mature research effort in its own right. A significant section of the research community is now shifting attention to certain problems which are peripheral, yet of immense significance to image retrieval systems, either directly or indirectly. Moreover, newly discovered problems are being solved with tools intended for image retrieval. In this section, we discuss such directions. Note that much of these peripheral ideas are in their infancy, and have likelihood of breaking into adulthood if sufficiently nurtured by the relevant research communities. Owing to the exploratory nature of the current approaches to these problems, a discussion is necessary on where these subfields are heading and what opportunities lie ahead for future innovation.

4.1. Words and Pictures

While contemplating problem of understanding picture content, it was soon learned that, in principle, associating those pictures with textual descriptions was only one step ahead. This led to the formulation of a new, but closely associated problem called *automatic image annotation*, often referred to as *auto-annotation* or *linguistic indexing*. The primary purpose of a practical content-based image retrieval system is to discover images pertaining to a given concept in the absence of reliable metadata. All attempts at automated concept discovery, annotation, or linguistic indexing essentially adhere to this objective. Annotation can facilitate image search through the use of text. If the resultant automated mapping between images and words can be trusted, text-based image searching can be semantically more meaningful than search in the absence of any text. Here we discuss two different schools of thought which have been used to address this problem.

4.1.1. Joint Word-Picture Modeling Approach. Many approaches to image annotation have been inspired by research in the text domain. Ideas from text modeling have been successfully imported to jointly model textual and visual data. In Duygulu et al. [2002], the problem of annotation is treated as a *translation* from a set of image segments to a set of words, in a way analogous to linguistic translation. A multimodal extension of a well-known hierarchical text model is proposed. Each word, describing a picture, is believed to have been generated by a node in a hierarchical concept tree. This assumption coheres with the hierarchical model for nouns and verbs adopted by Wordnet [Miller 1995]. This *translation model* is extended [Jin et al. 2005] to eliminate uncorrelated words from among those generated, making use of the Wordnet ontology. In Blei and Jordan [2003], the latent Dirichlet allocation (LDA) model is proposed for modeling associations between words and pictures.

In all such approaches, images are typically represented by properties of each of their segments, or *blobs*. Once all the pictures have been segmented, quantization can be used to obtain a finite vocabulary of blobs. Thus, the pictures under such models are treated as bags of words and blobs, each of which are assumed generated by *aspects*. Aspects are hidden variables which spawn a multivariate distribution over blobs and a multinomial distribution over words. Once the joint word-blob probabilities have been learned, the annotation problem for a given image is reduced to a likelihood problem relating blobs to words. The spatial relationships between blobs are not directly captured by the model. However, this is expected to be implicitly modeled in the generative distribution. Most of these techniques rely on precise segmentation, an issue which is still challenging. Despite the limitations, such modeling approaches remain popular.

Cross-media relevance models have been used for image annotation in Jeon et al. [2003] and Lavrenko et al. [2003]. A closely related approach involves coherent language models, and exploits word-to-word correlations to strengthen annotation decisions [Jin et al. 2004]. All the annotation strategies discussed so far model visual and textual features separately prior to association. A departure from this trend is seen in Monay and Gatica-Perez [2003], where probabilistic latent semantic analysis (PLSA) is used on uniform vectored data consisting of both visual features and textual annotations. This model is extended to a *nonlinear* latent semantic analysis for image annotation in Liu and Tang [2005].

4.1.2. Supervised Categorization Approach. An alternative approach is to treat image annotation as a supervised categorization problem. Concept detection through supervised classification, involving simple concepts such as city, landscape, and sunset, is achieved with high accuracy in Vailaya et al. [2001]. More recently, image annotation using both a novel structure-composition model and a WordNet-based word saliency measure has been proposed in Datta et al. [2007]. One of the earliest attempts at image annotation can be found in Li and Wang [2003]. The system, ALIP (automatic linguistic indexing of pictures) uses a 2D multiresolution hidden-Markov-models-based approach to capture inter- and intrascale spatial dependencies of image features of given semantic categories. Models for individual categories are learned independently and stored. The annotation step involves calculating likelihoods of the query image, given each learned model/category, and choosing annotations with bias toward statistically salient words corresponding to the most likely categories. A real-time image annotation system ALIPR (automatic linguistic indexing of pictures—real time) has been recently proposed in Li and Wang [2006]. ALIPR inherits its high-level learning architecture from ALIP. However, the modeling approach is simpler, hence leading to real-time computations of statistical likelihoods. Being the first real-time image annotation engine, ALIPR has generated considerable interest for real-world applications [Alipr 2006].

Learning concepts from user feedback in a dynamically changing image database using Gaussian mixture models is discussed in Dong and Bhanu [2003]. An approach to *soft* annotation, using Bayes point machines to give images a confidence level for each trained semantic label, is explored in Chang et al. [2003]. This vector of confidence labels can be exploited to rank relevant images in case of a keyword search. A confidence-based dynamic ensemble of SVM classifiers is used for annotation in Li et al. [2003]. Multiple-instance-learning-based approaches have been proposed for semantic categorization of images [Chen and Wang 2004] and to learn the correspondence between image regions and keywords [Yang et al. 2005a]. Concept learning based on a fusion of complementary classification techniques with limited training samples is proposed in Natsev et al. [2005]. Annotating images in dynamic settings (e.g., Yahoo! Flickr), where images and publicly generated tags arrive into a system asynchronously over time, has been explored using a metalearning framework in Datta et al. [2007].

Discussion. Automated annotation is widely recognized as an extremely difficult issue. We humans segment objects better than machines, having learned to associate over a long period of time, through multiple viewpoints, and literally through a "streaming video" at all times. This partly accounts for our natural segmentation capability. The association of words and blobs becomes truly meaningful only when blobs isolate objects well. Moreover, how exactly our brains make this association is unclear. While biology tries to answer this fundamental question, researchers in information retrieval tend to take a pragmatic stand in that they aim to build systems of practical significance. Ultimately, the desire is to be able to use keyword queries for all, images regardless of any manual annotations that they may have. To this end, a recent attempt at bridging the retrieval-annotation gap has been made [Datta et al. 2007].

4.2. Stories and Pictures

While the association between words and pictures is fairly well studied, deciding on an appropriate picture set for a given story is a relatively new problem. Attempts at tackling this problem are made in Barnard et al. [2003] and Joshi et al. [2006b]. By a story, we refer to a descriptive piece of text suitable for illustration in a practical sense. Possible applications of such systems could be automatic illustration of news articles at news agencies, or educational story illustration in textbooks.

The problem, however, poses several challenges: (1) People might attach different levels of importance to the ideas, concepts, and places discussed in a story. This subjectivity is hard to quantify and may be a result of past experiences, dislikes, and prejudices. (2) Any illustration system is constrained by the image repository from which the system selects pictures. An automated system may misperform if relevant pictures are not present or poorly represented in the repository. (3) Certain concepts might be overrepresented in the repository. Choosing a few representative pictures would then require a ranking scheme to discriminate among relevant pictures by some means. It is not easily perceived on what criteria this discrimination should be based.

A practical system which performs this task would require some way of identifying relevant keywords in a story and using a ranking scheme to determine representative pictures. In Barnard et al. [2003], the idea of auto-illustration is introduced as an inverse problem of auto-annotation. In Joshi et al. [2006b], image importance with respect to a story is quantified by the use of a mutual reinforcement principle. Given an annotated image database, pair-wise reinforcement is based on both visual similarity as well as Wordnet-based lexical similarity. This importance criteria are then used for choosing ideal pictures to illustrate the story in question.

Discussion. Evidently, work in this direction has been very limited, even though the problem is one of practical importance. One reason for this could be that the goals of auto-illustration or story-picturing are not as clearly defined as CBIR or image annotation. This brings us to the question of evaluation. How do we differentiate good illustrations from poor ones? The approach taken in Joshi et al. [2006b] is to exploit user studies to determine the agreement between human preference and automatic selection of pictures. Other, better approaches to evaluation may be possible. One thing is clear. however: A concrete formulation to the problem and an acceptable evaluation strategy for solutions are essentially two sides of the same coin.

4.3. Aesthetics and Pictures

Thus far, the focus of CBIR has been on semantics. There have been numerous discussions on the semantic gap. Imagine a situation where this gap has been bridged. This would mean, for example, finding all "dog" pictures in response to a "dog" query. In text-based search engines, a query containing "dog" will yield millions of Web pages. The smart search engine will then try to analyze the query to rank the best matches higher. The rationale for doing so is that of predicting what is most desirable based on the query. What, in CBIR, is analogous to such ranking, given that a large subset of the images are determined to be semantically relevant? This question has been recently addressed in Datta et al. [2006].

We conjecture that one way to distinguish among images of similar semantics is by their *quality*. Quality can be perceived at two levels: one involving concrete image parameters like size, aspect ratio, and color depth, and the other involving higher-level perception, which we denote as *aesthetics*. While it is trivial to rank images based on the former, the differences may not be significant enough to use as ranking criteria. On the other hand, aesthetics concerns the kind of emotions a picture arouses in people. Given this vague definition and the subjectivity associated with emotion, it is open to dispute how to aesthetically distinguish pictures. In our opinion, modeling aesthetics of images is an important open problem. Given a feasible model, a new dimension to image understanding will be added, benefiting CBIR and allied communities.

Discussion. The question remains as to how this problem can be approached. Given the high subjectivity of aesthetics, it may help to redefine the goal as a model that can characterize aesthetics *in general*. One way to model aesthetics in general is to study photo rating trends in public photo-sharing communities such as Photo.Net [1993], an approach that has been followed in Datta et al. [2006]. The site supports peer-rating of photographs based on aesthetics. This has generated a large database of ratings corresponding to the over one million photographs hosted. A discussion on the significance of these ratings, and aesthetic quality in general, can be found in the Photo.Net

RatingSystem page at PhotoNet [1993]. The idea of learning to assess visual aesthetics from such training data has been further pursued, for the purpose of selecting high-quality pictures and eliminating low-quality ones from image collections, in Datta et al. [2007]. Nonetheless, there *is* one caveat: Uncontrolled publicly collected data is naturally inclined to noise. When drawing conclusions about the data, this assumption must be kept in mind. Alternatively, ways to get around the noisy portions must be devised.

4.4. Art, Culture, and Pictures

Art and culture have always played an important role in human lives. Over the centuries, hundreds of museums and art galleries have preserved our diverse cultural heritage and served as important sources of education and learning. However, of late, concerns are being expressed to archive all ancient historical and cultural materials in digital form for posterity [Chen et al. 2005]. This is particularly important for two reasons, given next.

—Computers have become and will remain the primary medium for learning and education in the years to come. Hence, digital representation of cultural artifacts and pictures is bound to increase their popularity. Moreover, accessing digital archives is effortless and can practically be done from any corner of the world.

—As opposed to digital media, cultural artifacts and old paintings are subject to wear with time, and are also prone to disasters and vandalism [Chen et al. 2005].

In such a scenario, a key application of CBIR technology is to help preserve and analyze our history in digital media form. Growing research interest in the field is evident from the fact that in the year 2004, *IEEE Transactions on Image Processing* organized a special issue to the discuss state-of-the-art in image processing applications for cultural heritage [IEEE 2004]. The main focus of this issue is on modeling, retrieval, and authentication of cultural heritage images. Besides facilitating search and retrieval in large art/cultural image databases, statistical learning techniques have also been proposed to capture properties of the brush strokes of painters [Melzer et al. 1998; Sablatnig et al. 1998; Li and Wang 2004; Lyu et al. 2004; Berezhnoy et al. 2005; Johnson et al. 2008]. Such techniques can potentially be used to study similarities and differences among artists across countries, cultures, and time. Comprehensive surveys on latest advances in art imaging research can be found in Martinez et al. [2002], Maitre et al. [2001], Barni et al. [2005], and Chen et al. [2005].

Discussion. While it is impossible that automatic image analysis techniques can match the experience of art *connoisseurs*, they can definitely be used to complement human expertise. Statistical methods can sometime capture subtle characteristics of art which even a human eye can miss [Lyu et al. 2004].

4.5. Web and Pictures

The Web connects systems to systems, systems to people, and people with other people. Hosting a system on the Web is significantly different from hosting it in a private network or on a single machine. What makes things different in the Web medium is that we can no longer make assumptions about the users, nor their understanding of, way of interacting with, contributions to, and expectations from the system. Moreover, Web-based systems muster support of the masses only as long as they are useful to them. Without support, there is no meaning to such a system. This makes the creation of Web-based CBIR systems more challenging than the core questions of CBIR, aggravated further by the fact that multimedia searching is typically more complex than generic searching [Jansen et al. 2003]. Thankfully, the problem has recently received a lot of attention from the community, enough to have a survey dedicated specifically to it [Kherfi et al. 2004].

While we cannot make assumptions about generic Web-based CBIR systems, those designed while keeping in mind specific communities can afford some assumptions. Web-based CBIR services for copyright protection, tourism, entertainment, crime prevention, research, and education are some domain-specific possibilities, as reported in Kherfi et al. [2004]. One of the key tasks of Web image retrieval is crawling images. A smart Web crawler that attempts to associate captions with images to extract useful metadata in the crawling process is reported in Rowe [2002].

There have been many algorithms proposed for image search based on surrounding text, including those implemented in Google and Yahoo! image search. Here we discuss work that exploits image content in part or in full for retrieval. One of the earlier systems for Web-based CBIR, *iFind*, incorporated relevance feedback and was proposed in Zhang et al. [2000]. More recently, *Cortina*, a combined content- and metadata-based image search engine has been made public [Quack et al. 2004]. Other approaches to Web-based image retrieval include mutual reinforcement [Wang et al. 2004b], bootstrapping for annotation propagation [Feng et al. 2004], and nonparametric density estimation with application to an art image collection [Smolka et al. 2004]. Image grouping methods such as unsupervised clustering are extremely critical for heterogeneous repositories such as the Web (as discussed in Section 3.3), and this is explored in Wang et al. [2004a], Gao et al. [2005], Cai et al. [2004], and Jing et al. [2006]. More recently, rank fusion for Web image retrieval from multiple online picture forums has been proposed [Zhang et al. 2006]. Innovative interface designs for Web image search have been explored in Yee et al. [2003] and Li et al. [2004]. The SIMPLIcity system [Wang et al. 2001] has been incorporated into popular Websites such as Airliners.net [2005], Global Memory Net [2006], and Terragalleria [2001].

Discussion. The impact of CBIR can be best experienced through a Web-based image search service that gains popularity to the extent of its text-based counterparts. Unfortunately, at the time of writing this survey, this goal is elusive. Having said that, the significant progress in CBIR for the Web raises hopes for such systems in the coming years.

4.6. Security and Pictures

The interactions between CBIR and information security had been nonexistent until recently, when new perspectives emerged to strengthen the ties. Two such perspectives are human interactive proofs (HIPs), and the enforcement of copyright protection.

While on the one hand we are constantly pushing the frontiers of science to design intelligent systems that can imitate human capabilities, we cannot deny the inherent security risks associated with extremely smart computer programs. One such risk is when Web sites or public servers are attacked by malicious programs that request service on massive scale. Programs can be written to automatically consume large amount of Web resources or to bias results in online voting. HIPs, also known as CAPTCHAs, are a savior in these situations. These are interfaces designed to differentiate between humans and automated programs, based on responses to posed questions. The most common CAPTCHAs use distorted text, as seen in public Web sites such as Yahoo!, MSN, and PayPal. Recently, a number of OCR-based techniques have been proposed to break text-based CAPTCHAs [Mori and Malik 2003]. This has paved the way for natural-image-based CAPTCHAs, owing to the fact that CBIR is generally considered a much more difficult problem than OCR. The first formalization of image-based CAPTCHAs is found in Chew and Tygar [2004], where pictures chosen at random are displayed and questions asked, such as what does the picture contain, which picture is the odd one out conceptually, etc. A problem with this approach is the possibility that CBIR and concept learning techniques such as Barnard et al. [2003] and Li and Wang [2003] can be used to attack image-based CAPTCHAs. This will eventually lead to the same problem faced by text-based CAPTCHAs. To alleviate this problem, a CBIR system is used as a validation technique in order to distort images before being presented to users [Datta et al. 2005]. The distortions are chosen such that, probabilistically, CBIR systems find it difficult to grasp the image concepts and hence are unable to simulate human responses.

The second issue is image copy protection and forgery detection. Photographs taken by one person and posted online are often copied and passed on as someone else's artistry. Logos and trademarks of well-established organizations have often been duplicated by lesser-known firms, with or without minor modification, and with a clear intention to mislead patrons. While plagiarism of this nature is a world-wide phenomenon today, protection of the relevant copyrights is a very challenging task. The use of CBIR to help identify and possible enforce these copyrights is a relatively new field of study. In the case of exact copies, detecting them is trivial: Extraction and comparison of a simple file signature is sufficient. However, when changes to the pictures or logos are made, image similarity measures such as those employed in CBIR are necessary. The changes could be one or more of down-sampling, lowering of color depth, warping, shearing, cropping, decolorizing, palette shifting, changing contrast/brightness, image stamping, etc. The problem then becomes one of *near-duplicate detection*, in which case the similarity measures must be robust to these changes. Interest point detectors for generating localized image descriptors that are robust to such changes have been used for near-duplicate detection in Ke et al. [2004]. A part-based image similarity measure derived from the stochastic matching of attributed relational graphs is exploited for near-duplicate detection in Zhang and Chang [2004].

Discussion. Much security research is on the anticipation of possible attack strategies. While image-based CAPTCHA systems anticipate the use of CBIR for attacks, near-duplicate detectors anticipate possible image distortion methods that a copyright infringer may employ. Whether CBIR proves useful to security is yet to be seen, but dabbling with problems of this nature certainly helps CBIR grow as a field. For example, as noted in Zhang and Chang [2004], near-duplicate detection also finds application in weaving news stories across diverse video sources for news summarization. The generation of new ideas as offshoots of (or in the process of) solving other problems is the very essence of this section.

4.7. Machine Learning and Pictures

While more often than not, machine learning has been used to help solve the fundamental problem of image retrieval, there are instances where new and generic machine learning and data mining techniques have been developed in attempts to serve this purpose. The correspondence-LDA [Blei and Jordan 2003] model, proposed for joint word-image modeling, has since been applied to problems in bioinformatics [Zheng et al. 2006]. Probabilistic graphical models such as 2D multiresolution hidden Markov models [Li and Wang 2003] and cross-media relevance models [Jeon et al. 2003], though primarily used for image annotation applications, are contributions to machine learning research. Similarly, multiple-instance-learning research has benefited by work on image categorization [Chen and Wang 2004]. Active learning using SVMs was proposed for relevance feedback [Tong and Chang 2001] and helped to popularize active learning in other domains as well.

Automatic learning of a similarity metric or distance from ground-truth data has been explored for various tasks, such as clustering and classification. One way to achieve

Table III. A Qualitative Requirement Analysis of Various CBIR Offshoots and Applications

Applications and Offshoots	Similarity Measure	User Feedback	Machine learning	Visualization	Scalability
Automatic annotation	optional	optional	essential	optional	optional
Story illustration	essential	desirable	essential	desirable	desirable
Image-based CAPTCHA	essential	essential	optional	essential	essential
Copy detection	essential	desirable	optional	desirable	essential
Visual aesthetics	optional	desirable	essential	desirable	optional
Web image search	essential	optional	optional	essential	essential
Art image analysis	optional	desirable	essential	desirable	desirable

this is to learn a generalized Mahalanobis distance metric, such as the general-purpose methods proposed in Xing et al. [2003] and Bar-Hillel et al. [2005]. On the other hand, kernel-based learning of image similarity, using context information, with applications to image clustering was explored in Wu et al. [2005]. This method could potentially be used for more generic cases of metric learning when given side-information. In the use of a Mahalanobis metric for distance computation, an implicit assumption is that the underlying data distribution is Gaussian, which may not always be appropriate. An important work uses a principled approach to determine appropriate similarity metrics based on the nature of the underlying distributions, which is determined using ground-truth data [Sebe et al. 2000]. In a subsequent work, a boosting approach to learning a *boosted distance* measure (that is analogous to the weighted Euclidean norm) has been applied to stereo matching and video motion tracking [Yu et al. 2006], as well as classification/recognition tasks on popular datasets [Amores et al. 2006].

Discussion. In regard to recognizing pictures, even humans undergo a learning process. So, it is not surprising to see the synergy between machine learning and image retrieval when it comes to training computers to do the same. In fact, the challenges associated with learning from images have actually helped to push the scientific frontier in machine learning research in its own right.

4.8. Epilogue

While Sections 2 and 3 discussed techniques and real-world aspects of CBIR, in this section, we have described applications that employ these techniques. In Table III we present a qualitative requirement analysis of the various applications, involving a mapping from the aspects (i.e., techniques and features) to these applications. The entries are intended to be interpreted in the following manner.

—*Essential*. Essential aspects are those that are required in all scenarios.

—*Optional*. These are aspects that may or may not be critical, depending on the specific goals.

—*Desirable*. These are aspects that are *likely* to add value to the application in all cases.

The distinction between classifying an aspect as "optional" or "desirable" can be understood by the following examples. Scalability for automatic annotation is termed "optional" here because such an application can serve two purposes: (1) to be able to quickly tag a large number of pictures in a short time; and (2) to be able to produce accurate and consistent tags to pictures, or to refine existing noisy tags, perhaps as an offline process. Because of the compromise made in achieving these two goals, their scalability requirements may be different. As a second example, consider that in art image analysis, having an expert user involved in every step of the analysis is highly "desirable," unlike in large-scale image annotation where a user validation at each step may be infeasible.

5. EVALUATION STRATEGIES

Whenever there are multiple competing products in the market, customers typically resort to statistics, reviews, and public opinions in order to make a well-informed selection. A direct analogy can be drawn for CBIR. With the numerous competing techniques and systems proposed and in operation, evaluation becomes a critical issue. Even from the point-of-view of researchers, a benchmark for the evaluation of CBIR would allow choosing from many different proposed ideas and to test new approaches against older ones. For any information retrieval system, a strategy for evaluation involves determining the following aspects.

—*An Appropriate Dataset for Evaluation*. The dataset should be general enough to cover a large range of semantics from a human point-of-view. Also, the dataset should be large enough for the evaluation to be statistically significant.

—*A Ground Truth for Relevance for the Problem at Hand*. Ground truth is a very subjective issue, especially for multimedia. Usually, people associate a given picture with a wide range of high-level semantics.

—*An Appropriate Metric and Criteria for Evaluating Competing Approaches*. The evaluation criteria should try to model human requirements from a population perspective.

Moreover, it is desirable to have a forum or gathering at regular intervals for discussing different approaches, as well as their respective performance and shortcomings using the evaluation strategy. The problem of CBIR evaluation, however, is very challenging. The aforementioned points often make it very difficult to decide upon an evaluation dataset and to obtain reliable ground truth for it. Deciding on a metric and evaluation criteria is another difficult problem. An objective evaluation of results could be unfair and incomplete, since CBIR technology is eventually expected to satisfy the needs of people who use it. In spite of these challenges, researchers have agreed upon certain evaluation datasets, benchmarks, and forums for multimedia retrieval evaluation. These are described as follows.

5.1. Evaluation Metrics

CBIR is essentially an information retrieval problem. Therefore, evaluation metrics have been quite naturally adopted from information retrieval research. Two of the most popular evaluation measures are described next.

—*Precision*. This refers to the percentage of retrieved pictures that are relevant to the query.

—*Recall*. This pertains to the percentage of all the relevant pictures in the search database which are retrieved.

Notice that when the query in question is a picture, relevance is extremely subjective. Information retrieval research has shown that precision and recall follow an inverse relationship. Precision falls while recall increases as the number of retrieved pictures, often termed as *scope*, increases. Hence, it is typical to have a high numeric value for both precision and recall. Traditionally, results are summarized as *precision-recall* curves or *precision-scope* curves. A criticism for precision stems from the fact that it is calculated for the entire retrieved set and unaffected by the respective rankings of relevant entities in the retrieved list.

A measure which addresses the aforesaid problem and that is very popular in CBIR community is *average precision* (AP). In a ranked list of retrieved entities with respect to a query, if precision is calculated at the depth of every relevant entity obtained, then average precision is given as the mean of all the individual precision scores. Obviously, this metric is highly influenced by high-ranked relevant entities, and not so much by those toward the bottom of the ranked list. The arithmetic mean of average precision calculated over a number of different queries is often reported as mean average precision (MAP) and is one of the evaluation measures used by the TRECVID community [TRECVID 2001]. A comprehensive overview and discussion on performance measures for CBIR has been presented in Huijsmans and Sebe [2005]. The authors of the cited work discuss the influence of individual class sizes to these measures, in a CBIR system. The importance of normalization of performance measures with respect to scope and to class sizes has been emphasized.

5.2. Evaluation Criteria

As observed in Shirahatti and Barnard [2005], CBIR is meaningful only in its service to human users. At the same time, it is difficult to quantify user requirements as objective relevance-based scores. As discussed in Section 2, users may be classified into several types based on their clarity of intent and search patterns. Depending upon the end-goal, a user may value different features of a CBIR system.

An interesting user-driven evaluation criteria has been proposed in Shirahatti and Barnard [2005]. The authors construct a mapping of various retrieval algorithm scores to human assessment of similarity. As a consequence, different retrieval algorithms can be evaluated against the same user-determined scale. Another work studies user information needs with respect to image retrieval, using American memory photo archives [Choi and Rasmussen 2002]. It has been observed that users of an image retrieval system value several important criteria such as image quality, clarity, and associated metadata, besides image semantics.

5.3. Evaluation Datasets and Forums

Traditionally, in the absence of benchmarks, Corel Stock Photos and Caltech101 [Caltech101 2004] have been used for CBIR evaluation. The authors of Caltech101 have released a new version of their dataset, called Caltech256, including 256 picture categories. The pitfalls of using Corel pictures have been discussed in Muller et al. [2002], and a more rigorous CBIR benchmarking is suggested therein. The Benchathlon Project [Benchathlon 2005; Gunther and Beratta 2001] was initiated to get the CBIR community to come together for formulating evaluation strategies. ImageCLEF [ImageCLEF 2006], a track as part of a cross-language evaluation forum, focuses on evaluation strategies for CBIR. Another important effort in this direction is the ImagEVAL workshop [ImagEVAL 2005] where the importance of user-oriented evaluation has been emphasized. The ImagEVAL effort stresses criteria such as the quality of user interface, the response time, and adaptiveness of a CBIR system to a new domain. The TRECVID benchmark is very popular in the CBIR community to validate their search and retrieval algorithms [TRECVID 2001; Smeaton and Over 2003]. The TRECVID workshop, conducted yearly by the National Institute of Science and Technology (NIST), attracts research teams from all over the world for addressing competitive problems in content-based video search and retrieval. A comprehensive overview of benchmarking in CBIR can be found in Muller et al. [2001].

5.4. Discussion

From the current trends and the effort being put into benchmarking in CBIR, the following design goals emerge.

—*Coverage*. Benchmarks should ideally cover the entire spectrum of cases expected in real-world scenarios. This should affect the choice of evaluation datasets.

—*Unbiasedness*. Benchmarks should not show any bias toward particular algorithms or methodologies. In particular, factors such as accuracy, speed, compatibility, and adaptiveness should be given as much importance as required for the target application.

—*User Focus*. General-purpose CBIR applications are designed for use by human users. A fair benchmark for such applications should adequately reflect user interest and satisfaction.

Evaluation is critical for both CBIR as well as its offshoot research areas. Ideally, evaluation should be subjective, context-specific, and community-based. For example, Web-based image retrieval is best judged by a typical sampling of Internet users, whereas evaluation of retrieval for biomedical applications will require users with domain knowledge and expertise. Automated annotation is best evaluated in the context of what detail the systems are aiming at. Depending on application, it may or may not be sufficient to label a rose as a flower. Illustration of stories can be best appreciated by how readers receive them.

In summary, evaluation is a vital component of system design that needs to be performed while keeping in mind the end-users. CBIR and its offshoots are no exceptions. Developing user-centric benchmarks is a next-generation challenge for researchers in CBIR and associated areas. However, it is important to maintain a balance between exploring new and exciting research problems and developing rigorous evaluation methods for the existing ones [Wang et al. 2006].

6. DISCUSSION AND CONCLUSIONS

We have presented a comprehensive survey highlighting current progress, emerging directions, the spawning of new fields, and methods for evaluation relevant to the young and exciting field of image retrieval. We have contrasted the early years of image retrieval with progress in the field in the current decade, and conjectured specific future directions alongside. We believe that the field will experience a paradigm shift in the foreseeable future, with the focus being more on application-oriented, domain-specific work, generating considerable impact in day-to-day life.

As part of an effort to better understand the field of image retrieval, we compiled research trends in image retrieval using Google Scholar's search tool and its computed citation scores. Graphs for publication counts and citation scores have been generated for: (1) subfields of image retrieval, and (2) venues/journals relevant to image retrieval research. Further analysis has been made on the impact of image retrieval on merging interests among different fields of study, such as multimedia (MM), machine learning (ML), information retrieval (IR), computer vision (CV), and human-computer interaction (HCI). Firstly, the trends indicate that the field is extremely diverse, and can only grow to be more so in the future. Second, we note that image retrieval has likely been the catalyst in the closer association, of hitherto unrelated fields of research. Third, interesting facts have emerged, such as: Most of the MM, CV, and AI work related to image retrieval has been published in information-related venues and received high citations. At the same time, AI-related work published in CV venues has generated considerable impact. At a higher level, the trends indicate that while aspects including systems, feature extraction, and relevance feedback have received a lot of attention, application-oriented aspects such as interface, visualization, scalability, and evaluation have traditionally received lesser consideration. We feel that for all practical purposes, these latter aspects should also be considered as equally important. Due to the dynamic nature of this information, we have decided to host it externally, and to update it from time to time, at `http://wang.ist.psu.edu/ survey/analysis`.

The quality (resolution and color depth), nature (dimensionality), and throughput (rate of generation) of the images acquired have all been on an upward growth path in recent times. With the advent of very large-scale images (e.g., Google and Yahoo! aerial maps), biomedical and astronomical imagery have become typically of high resolution/dimension and are often captured at high throughput, posing yet new challenges to image retrieval research. A long-term goal of research should therefore also include the ability to make high-resolution, high-dimension, and high-throughput images searchable by content. Meanwhile, we do hope that the quest for robust and reliable image understanding technology will continue. The future of CBIR depends a lot on the collective focus and overall progress in each aspect of image retrieval, and how much the average individual stands to benefit from it.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their constructive comments. Hanan Samet offered helpful suggestions. J. Z. Wang gratefully acknowledges Takeo Kanade for his encouragement.

REFERENCES

AIGRAIN, P., ZHANG, H., AND PETKOVIC, D. 1996. Content-based representation and retrieval of visual media: A review of the state-of-the-art. *Multimed. Tools Appl. 3*, 3, 179–202.

AIRLINERS.NET. 2005. Airliners.net homepage. http://www.airliners.net.

ALIPR. 2006. Alipr homepage. http://www.alipr.com.

AMORES, J., SEBE, N., AND RADEVA, P. 2005. Fast spatial pattern discovery integrating boosting with constellations of contextual descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

AMORES, J., SEBE, N., AND RADEVA, P. 2006. Boosting the distance estimation: Application to the k-nearest neighbor classifier. *Pattern Recogn. Lett. 27*, 3, 201–209.

AMORES, J., SEBE, N., RADEVA, P., GEVERS, T., AND SMEULDERS, A. 2004. Boosting contextual information in content-based image retrieval. In *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR) at the International Multimedia Conference*.

ARMITAGE, L. H. AND ENSER, P. G. B. 1997. Analysis of user need in image archives. *J. Inf. Sci. 23*, 4, 287–299.

ASSFALG, J., DEL BIMBO, A., AND PALA, P. 2002. Three-Dimensional interfaces for querying by example in content-based image retrieval. *IEEE Trans. Visualiz. Comput. Graphics 8*, 4, 305–318.

BAR-HILLEL, A., HERTZ, T., SHENTAL, N., AND WEINSHALL, D. 2005. Learning a Mahalanobis metric from equivalence constraints. *J. Mach. Learn. Res. 6*, 937–965.

BARNARD, K., DUYGULU, P., FORSYTH, D., DE FREITAS, N., BLEI, D. M., AND JORDAN, M. I. 2003. Matching words and pictures. *J. Mach. Learn. Res. 3*, 1107–1135.

BARNI, M., PELAGOTTI, A., AND PIVA, A. 2005. Image processing for the analysis and conservation of paintings: Opportunities and challenges. *IEEE Signal Process. Mag. 22*, 141–144.

BARTOLINI, I., CIACCIA, P., AND PATELLA, M. 2005. Warp: Accurate retrieval of shapes using phase of Fourier descriptors and time warping distance. *IEEE Trans. Pattern Anal. Mach. Intell. 27*, 1, 142–147.

BELONGIE, S., MALIK, J., AND PUZICHA, J. 2002. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 4, 509–522.

BENCHATHLON. 2005. Benchathlon homepage. http://www.benchathlon.net.

BEREZHNOY, I. E., POSTMA, E. O., AND HERIK, J. V. D. 2005. Computerized visual analysis of paintings. In *Proceedings of the International Conference on the Association for History and Computing*.

BERRETTI, S., BIMBO, A. D., AND VICARIO, E. 2001. Efficient matching and indexing of graph models in content-based retrieval. *IEEE Trans. Pattern Anal. Mach. Intell. 23*, 10, 1089–1105.

BERRETTI, S. AND DEL BIMBO, A. 2006. Modeling spatial relationships between 3D objects. In *Proceedings of the 18th International IEEE Conference on Pattern Recognition (ICPR)*.

BERRETTI, S., DEL BIMBO, A., AND PALA, P. 2000. Retrieval by shape similarity with perceptual distance and effective indexing. *IEEE Trans. Multimed. 2*, 4, 225–239.

BERRETTI, S., DEL BIMBO, A., AND VICARIO, E. 2003. Weighted walkthroughs between extended entities for retrieval by spatial arrangement. *IEEE Trans. Multimed. 5*, 1, 52–70.

BERTINI, E., CALI, A., CATARCI, T., GABRIELLI, S., AND KIMANI, S. 2005. Interaction-Based adaptation for small screen devices. In *Proceedings of the 10th International Conference on User Modeling (UM)*. Lecture Notes in Computer Science, Vol. 3538. Springer, 277–281.

BERTINI, M., CUCCHIARA, R., DEL BIMBO, A., AND PRATI, A. 2003. Object and event detection for semantic annotation and transcoding. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*.

BLEI, D. M. AND JORDAN, M. I. 2003. Modeling annotated data. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

BOHM, C., BERCHTOLD, S., AND KEIM, D. A. 2001. Searching in high-dimensional space index structures for improving the performance of multimedia databases. *ACM Comput. Surv. 33*, 3, 322–373.

BOUCHARD, G. AND TRIGGS, B. 2005. Hierarchical part-based visual object categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

CAI, D., HE, X., LI, Z., MA, W. Y., AND WEN, J. R. 2004. Hierarchical clustering of www image search results using visual, textual and link information. In *Proceedings of the ACM International Conference on Multimedia*.

CALTECH101. 2004. http://www.vision.caltech.edu/image_datasets/caltech101/caltech101.html.

CARBALLIDO-GAMIO, J., BELONGIE, S., AND MAJUMDAR, S. 2004. Normalized cuts in 3-D for spinal MRI segmentation. *IEEE Trans. Medical Imag. 23*, 1, 36–44.

CARNEIRO, G. AND LOWE, D. 2006. Sparse flexible models of local features. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

CARNEIRO, G. AND VASCONCELOS, N. 2005. Minimum Bayes error features for visual recognition by sequential feature selection and extraction. In *Proceedings of the Canadian Conference on Computer and Robot Vision*.

CARSON, C., BELONGIE, S., GREENSPAN, H., AND MALIK, J. 2002. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 8, 1026–1038.

CHALECHALE, A., NAGHDY, G., AND MERTINS, A. 2005. Sketch-based image matching using angular partitioning. *IEEE Trans. Syst. Man Cybernet. 35*, 1, 28–41.

CHANG, E. Y., GOH, K., SYCHAY, G., AND WU, G. 2003. CBSA: Content-based soft annotation for multimodal image retrieval using Bayes point machines. *IEEE Trans. Circ. Systems Video Technol. 13*, 1, 26–38.

CHANG, S., SHI, Q., AND YAN, C. 1987. Iconic indexing by 2-D strings. *IEEE Trans. Pattern Anal. Mach. Intell. 9*, 3, 413–427.

CHANG, S., YAN, C., DIMITROFF, D., AND ARNDT, T. 1988. An intelligent image database system. *IEEE Trans. Softw. Eng. 14*, 5, 681–688.

CHANG, S.-F., SMITH, J., BEIGI, M., AND BENITEZ, A. 1997. Visual information retrieval from large distributed online repositories. *Commun. ACM 40*, 12, 63–71.

CHEN, C.-C., WACTLAR, H., WANG, J. Z., AND KIERNAN, K. 2005. Digital imagery for significant cultural and historical materials—An emerging research field bridging people, culture, and technologies. *Int. J. Digital Libr. 5*, 4, 275–286.

CHEN, J., PAPPAS, T., MOJSILOVIC, A., AND ROGOWITZ, B. 2002. Adaptive image segmentation based on color and texture. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*.

CHEN, L. Q., XIE, X., FAN, X., MA, W. Y., ZHANG, H. J., AND ZHOU, H. Q. 2003. A visual attention model for adapting images on small displays. *Multimed. Syst. 9*, 4, 353–364.

CHEN, Y. AND WANG, J. Z. 2002. A region-based fuzzy feature matching approach to content-based image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 9, 252–1267.

CHEN, Y. AND WANG, J. Z. 2004. Image categorization by learning and reasoning with regions. *J. Mach. Learn. Res. 5*, 913–939.

CHEN, Y., WANG, J. Z., AND KROVETZ, R. 2005. Clue: Cluster-based retrieval of images by unsupervised learning. *IEEE Trans. Image Process. 14*, 8, 1187–1201.

CHEN, Y., ZHOU, X., AND HUANG, T. S. 2002. One-class SVM for learning in image retrieval. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*.

CHEW, M. AND TYGAR, J. D. 2004. Image recognition captchas. In *Proceedings of the Information Security Conference.*

CHOI, Y. AND RASMUSSEN, E. M. 2002. User's relevance criteria in image retrieval in American history. *Inf. Process. Manage. 38*, 5, 695–726.

CHRISTEL, M. G. AND CONESCU, R. M. 2005. Addressing the challenge of visual information access from digital image and video libraries. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*.

CIACCIA, P., PATELLA, M., AND ZEZULA, P. 1997. M-Tree: An efficient access method for similarity search in metric spaces. In *Proceedings of the International Conference on Very Large Databases (VLDB)*.

CNN. 2005. Computer decodes Mona Lisa's smile. http://www.cnn.com/2005/TECH/12/16/mona.lisa.smile/index.html.

COMANICIU, D. AND MEER, P. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 5, 603–619.

COX, I. J., MILLER, M. L., MINKA, T. P., PAPATHOMAS, T. V., AND YIANILOS, P. N. 2000. The Bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments. *IEEE Trans. Image Process. 9*, 1, 20–37.

CSILLAGHY, A., HINTERBERGER, H., AND BENZ, A. 2000. Content based image retrieval in astronomy. *Inf. Retriev. 3*, 3, 229–241.

CUCCHIARA, R., GRANA., C., AND PRATI, A. 2003. Semantic video transcoding using classes of relevance. *Int. J. Image Graph. 3*, 1, 145–170.

CUNNINGHAM, S. J., BAINBRIDGE, D., AND MASOODIAN, M. 2004. How people describe their image information needs: A grounded theory analysis of visual arts queries. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*.

DAGLI, C. AND HUANG, T. S. 2004. A framework for grid-based image retrieval. In *Proceedings of the International IEEE Conference on Pattern Recognition (ICPR)*.

DATTA, R., GE, W., LI, J., AND WANG, J. Z. 2007. Toward bridging the annotation-retrieval gap in image search. *IEEE Multimed. 14*, 3, 24–35.

DATTA, R., JOSHI, D., LI, J., AND WANG, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

DATTA, R., JOSHI, D., LI, J., AND WANG, J. Z. 2007. Tagging over time: Real-world image annotation by lightweight meta-learning. In *Proceedings of the ACM International Conference on Multimedia*.

DATTA, R., LI, J., AND WANG, J. Z. 2005. IMAGINATION: A robust image-based captcha generation system. In *Proceedings of the ACM International Conference on Multimedia*.

DATTA, R., LI, J., AND WANG, J. Z. 2007. Learning the consensus on visual quality for next-generation image management. In *Proceedings of the ACM International Conference on Multimedia*.

DE SILVA, V. AND TENENBAUM, J. 2003. Global versus local methods in nonlinear dimensionality reduction. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*.

DEL BIMBO, A. D. 1999. *Visual Information Retrieval*. Morgan Kaufmann.

DEL BIMBO, A. AND PALA, P. 1997. Visual image retrieval by elastic matching of user sketches. *IEEE Trans. Pattern Anal. Mach. Intell. 19*, 2, 121–132.

DENG, Y. AND MANJUNATH, B. 2001. Unsupervised segmentation of color-texture regions in images and video. *IEEE Trans. Pattern Anal. Mach. Intell. 23*, 8, 800–810.

DENG, Y., MANJUNATH, B. S., KENNEY, C., MOORE, M. S., AND SHIN, H. 2001. An efficient color representation for image retrieval. *IEEE Trans. Image Process. 10*, 1, 140–147.

DO, M. N. AND VETTERLI, M. 2002. Wavelet-Based texture retrieval using generalized Gaussian density and Kullback-Leibler distance. *IEEE Trans. Image Process. 11*, 2, 146–158.

DONG, A. AND BHANU, B. 2003. Active concept learning for image retrieval in dynamic databases. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

DU, Y. AND WANG, J. Z. 2001. A scalable integrated region-based image retrieval system. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*.

DUYGULU, P., BARNARD, K., DE FREITAS, N., AND FORSYTH, D. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

FAGIN, R. 1997. Combining fuzzy information from multiple systems. In *Proceedings of the ACM-SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*.

FANG, Y. AND GEMAN, D. 2005. Experiments in mental face retrieval. In *Proceedings of the International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*.

FANG, Y., GEMAN, D., AND BOUJEMAA, N. 2005. An interactive system for mental face retrieval. In *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR) at the International Multimedia Conference*.

FENG, H., SHI, R., AND CHUA, T. S. 2004. A bootstrapping framework for annotating and retrieving www images. In *Proceedings of the ACM International Conference on Multimedia*.

FERGUS, R., PERONA, P., AND ZISSERMAN, A. 2003. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

FERGUS, R., PERONA, P., AND ZISSERMAN, A. 2005. A sparse object category model for efficient learning and exhaustive recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

FINLAYSON, G. 1996. Color in perspective. *IEEE Trans. Pattern Anal. Mach. Intell. 18*, 10, 1034–1038.

FLICKNER, M., SAWHNEY, H., NIBLACK, W., ASHLEY, J., HUANG, Q., 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. 28*, 9, 23–32.

FLICKR. 2002. The flickr homepage. http://www.flickr.com.

GAO, B., LIU, T.-Y., QIN, T., ZHENG, X., CHENG, Q.-S., AND MA, W.-Y. 2005. Web image clustering by consistent utilization of visual features and surrounding texts. In *Proceedings of the ACM International Conference on Multimedia*.

GEVERS, T. AND SMEULDERS, A. 2000. Pictoseek: Combining color and shape invariant features for image retrieval. *IEEE Trans. Image Process, 9*, 1, 102–119.

GLOBALMEMORYNET. 2006. Global Memory Net homepage. http://www.memorynet.org.

GOH, K.-S., CHANG, E. Y., AND CHENG, K.-T. 2001. SVM binary classifier ensembles for image classification. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*.

GOH, K.-S., CHANG, E. Y., AND LAI, W.-C. 2004. Multimodal concept-dependent active learning for image retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

GOOGLE SCHOLAR. 2004. Google scholar homepage. http://scholar.google.com.

GORDON, S., GREENSPAN, H., AND GOLDBERGER, J. 2003. Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

GOUET, V. AND BOUJEMAA, N. 2002. On the robustness of color points of interest for image retrieval. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*.

GRAUMAN, K. AND DARRELL, T. 2005. Efficient image matching with distributions of local invariant features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

GUNTHER, N. J. AND BERATTA, G. B. 2001. Benchmark for image retrieval using distributed systems over the Internet: Birds-I. In *Proceedings of the SPIE Conference on Internet Imaging III, Vol. 4311*, 252–267.

GUPTA, A. AND JAIN, R. 1997. Visual information retrieval. *Commun. ACM 40*, 5, 70–79.

GUYON, I. AND ELISSEEFF, A. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res. 3*, 1157–1182.

HADJIDEMETRIOU, E., GROSSBERG, M. D., AND NAYAR, S. K. 2004. Multiresolution histograms and their use for recognition. *IEEE Trans. Pattern Anal. Mach. Intell. 26*, 7, 831–847.

HAN, J., NGAN, K. N., LI, M., AND ZHANG, H.-J. 2005. A memory learning framework for effective image retrieval. *IEEE Trans. Image Process. 14*, 4, 511–524.

HARALICK, R. 1979. Statistical and structural approaches to texture. *Proc. IEEE 67*, 5, 786–804.

HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. 2001. *The Elements of Statistical Learning*. Springer-Verlag.

HAUPTMANN, A. G. AND CHRISTEL, M. G. 2004. Successful approaches in the TREC video retrieval evaluations. In *Proceedings of the ACM International Conference on Multimedia*.

HE, J., LI, M., ZHANG, H.-J., TONG, H., AND ZHANG, C. 2004a. Manifold-ranking based image retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

HE, J., LI, M., ZHANG, H.-J., TONG, H., AND ZHANG, C. 2004b. Mean version space: A new active learning method for content-based image retrieval. In *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR) at the International Multimedia Conference*.

HE, X. 2004. Incremental semi-supervised subspace learning for image retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

HE, X., MA, W.-Y., AND ZHANG, H.-J. 2004c. Learning an image manifold for retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

HOI, C.-H. AND LYU, M. R. 2004a. Group-based relevance feedback with support vector machine ensembles. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*.

Hoi, C.-H. and Lyu, M. R. 2004b. A novel log-based relevance feedback technique in content-based image retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

Hoiem, D., Sukthankar, R., Schneiderman, H., and Huston, L. 2004. Object-based image retrieval using the statistical structure of images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, J., Ravi Kumar, S., Mitra, M., Zhu, W.-J., and Zabih, R. 1999. Spatial color indexing and applications. *Int. J. Comput. Vision 35*, 3, 245–268.

Huijsmans, D. P. and Sebe, N. 2005. How to complete performance graphs in content-based image retrieval: Add generality and normalize scope. *IEEE Trans. Pattern Anal. Mach. Intell. 27*, 2, 245–251.

Huynh, D. F., Drucker, S. M., Baudisch, P., and Wong, C. 2005. Time quilt: Scaling up zoomable photo browsers for large, unstructured photo collections. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

IEEE(TIP). 2004. Special issue on image processing for cultural heritage. *IEEE Trans. Image Process. 13*, 3.

ImageCLEF. 2006. The CLEF cross language image retrieval track (ImageCLEF) homepage. http://ir.shef.ac.uk/imageclef.

ImagEVAL. 2005. The ImageEval homepage. http://www.imageval.org.

Iqbal, Q. and Aggarwal, J. K. 2002. Retrieval by classification of images containing large manmade objects using perceptual grouping. *Pattern Recogn. J. 35*, 7, 1463–1479.

Jaimes, A., Omura, K., Nagamine, T., and Hirata, K. 2004. Memory cues for meeting video retrieval. In *Proceedings of the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE) at the ACM International Multimedia Conference*.

Jaimes, A., Sebe, N., and Gatica-Perez, D. 2006. Human-centered computing: A multimedia perspective. In *Proceedings of the ACM International Conference on Multimedia (Special Session on Human-Centered Multimedia)*.

Jain, A. and Farrokhnia, F. 1990. Unsupervised texture segmentation using Gabor filters. *Proceedings of the International Conference on Systems, Man and Cybernetics*.

Jansen, B. J., Spink, A., and Pedersen, J. 2003. An analysis of multimedia searching on AltaVista. In *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR) at the International Multimedia Conference*.

Jeon, J., Lavrenko, V., and Manmatha, R. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jeong, S., Won, C. S., and Gray, R. 2004. Image retrieval using color histograms generated by Gauss mixture vector quantization. *Comput. Vision Image Understand. 9*, 1–3, 44–66.

Johnson Jr., C. R., Hendriks, E., Berezhony, I., Brevdo, E., Hughes, S., Daubechies, I., Li, J., Postma, E., and Wang, J. Z. 2008. Image processing for artist identification—computerized analysis of Vincent van Gogh's painting brush stokes. *IEEE Sign. Process. 25* (Special Issue on Visual Cultural Heritage).

Jin, R., Chai, J. Y., and Si, L. 2004. Effective automatic image annotation via a coherent language model and active learning. In *Proceedings of the ACM International Conference on Multimedia*.

Jin, R. and Hauptmann, A. 2002. Using a probabilistic source model for comparing images. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*.

Jin, Y., Khan, L., Wang, L., and Awad, M. 2005. Image annotations by combining multiple evidence and Wordnet. In *Proceedings of the ACM International Conference on Multimedia*.

Jing, F., Li, M., Zhang, H.-J., and Zhang, B. 2004a. An efficient and effective region-based image retrieval framework. *IEEE Trans. Image Process. 13*, 5, 699–709.

Jing, F., Li, M., Zhang, H.-J., and Zhang, B. 2004b. Relevance feedback in region-based image retrieval. *IEEE Trans. Circ. Syst. Video Technol. 14*, 5, 672–681.

Jing, F., Li, M., Zhang, H. J., and Zhang, B. 2005. A unified framework for image retrieval using keyword and visual features. *IEEE Trans. Image Process. 14*, 7, 978–989.

Jing, F., Wang, C., Yao, Y., Deng, K., Zhang, L., and Ma, W. Y. 2006. Igroup: Web image search results clustering. In *Proceedings of the ACM International Conference on Multimedia*.

Joshi, D., Datta, R., Zhuang, Z., Weiss, W., Friedenberg, M., Wang, J., and Li, J. 2006a. Paragrab: A comprehensive architecture for Web image management and multimodal querying. In *Proceedings of the International Conference on VLDB.*

Joshi, D., Naphade, M., and Natsev, A. 2007. A greedy performance driven algorithm for decision fusion learning. In *Proceedings of the IEEE International Conference on Image Processing (ICIP).*

JOSHI, D., WANG, J. Z., AND LI, J. 2006b. The story picturing engine—a system for automatic text illustration. *ACM Trans. Multimed. Comput. Commun. Appl. 2*, 1, 68–89.

KASTER, T., PFEIFFER, M., AND BAUCKHAGE, C. 2003. Combining speech and haptics for intuitive and efficient navigation through image databases. In *Proceedings of the 5th International Conference on Multimidia Interfaces (ICMI)*.

KE, Y., SUKTHANKAR, R., AND HUSTON, L. 2004. Efficient near-duplicate detection and subimage retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

KHERFI, M. L., ZIOU, D., AND BERNARDI, A. 2004. Image retrieval from the World Wide Web: Issues, techniques, and systems. *ACM Comput. Surv. 36*, 1, 35–67.

KIM, D.-H. AND CHUNG, C.-W. 2003. Qcluster: Relevance feedback using adaptive clustering for content based image retrieval. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*.

KIM, Y. S., STREET, W. N., AND MENCZER, F. 2000. Feature selection in unsupervised learning via evolutionary search. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*.

KO, B. AND BYUN, H. 2002. Integrated region-based image retrieval using region's spatial relationships. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*.

KOTOULAS, L. AND ANDREADIS, I. 2003. Colour histogram content-based image retrieval and hardware implementation. *IEEE Proc. Circ. Dev. Syst. 150*, 5, 387–393.

LAAKSONEN, J., KOSKELA, M., LAAKSO, S., AND OJA, E. 2001. Self-organizing maps as a relevance feedback technique in content-based image retrieval. *Pattern Anal. Appl. 4*, 140–152.

LAAKSONEN, J., KOSKELA, M., AND OJA, E. 2002. PicSom-Self-Organizing image retrieval with mpeg-7 content descriptors. *IEEE Trans. Neural Netw. 13*, 4, 841–853.

LATECKI, L. J. AND LAKAMPER, R. 2000. Shape similarity measure based on correspondence of visual parts. *IEEE Trans. Pattern Anal. Mach. Intell. 22*, 10, 1185–1190.

LAVRENKO, V., MANMATHA, R., AND JEON, J. 2003. A model for learning the semantics of pictures. In *Proceedings of the Conference on Advances in Neural Information Pracessing Systems (NIPS)*.

LAZEBNIK, S., SCHMID, C., AND PONCE, J. 2003. Affine-invariant local descriptors and neighborhood statistics for texture recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

LEVINA, E. AND BICKEL, P. 2001. The earth mover's distance is the Mallows distance: Some insights from statistics. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

LEW, M., SEBE, N., DJERABA, C., AND JAIN, R. 2006. Content-based multimedia information retrieval: State-of-the-art and challenges. *ACM Trans. Multimed. Comput. Commun. Appl. 2*, 1, 1–19.

LI, B., GOH, K.-S., AND CHANG, E. Y. 2003. Confidence-based dynamic ensemble for image annotation and semantics discovery. In *Proceedings of the ACM International Conference on Multimedia*.

LI, J. 2005. Two-scale image retrieval with significant meta-information feedback. In *Proceedings of the ACM International Conference on Multimedia*.

LI, J., GRAY, R. M., AND OLSHEN, R. A. 2000. Multiresolution image classification by hierarchical modeling with two dimensional hidden Markov models. *IEEE Trans. Inf. Theory 46*, 5, 1826–1841.

LI, J., NAJMI, A., AND GRAY, R. M. 2000. Image classification by a two dimensional hidden Markov model. *IEEE Trans. Signal Process. 48*, 2, 527–533.

LI, J. AND SUN, H.-H. 2003. On interactive browsing of large images. *IEEE Trans. Multimed. 5*, 4, 581–590.

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

LI, J. AND WANG, J. Z. 2004. Studying digital imagery of ancient paintings by mixtures of stochastic models. *IEEE Trans. Image Process. 13*, 3, 340–353.

LI, J. AND WANG, J. Z. 2006. Real-Time computerized annotation of pictures. In *Proceedings of the ACM International Conference on Multimedia*.

LI, J. AND WANG, J. Z. 2008. Real-time computerized annotation of pictures. *IEEE Trans. Pattern Anal. Mach. Intell. 30*, 6, 985–1002.

LI, J., WANG, J. Z., AND WIEDERHOLD, G. 2000. IRM: Integrated region matching for image retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

LI, Z.-W., XIE, X., LIU, H., TANG, X., LI, M., AND MA, W.-Y. 2004. Intuitive and effective interfaces for www image search engines. In *Proceedings of the ACM International Conference on Multimedia*.

LIN, Y.-Y., LIU, T.-L., AND CHEN, H.-T. 2005. Semantic manifold learning for image retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

LIU, W. AND TANG, X. 2005. Learning an image-word embedding for image auto-annotation on the nonlinear latent space. In *Proceedings of the ACM International Conference on Multimedia*.

LU, Y., HU, C., ZHU, X., ZHANG, H., AND YANG, Q. 2000. A unified framework for semantics and feature based relevance feedback in image retrieval systems. In *Proceedings of the ACM International Conference on Multimedia*.

LYU, S., ROCKMORE, D., AND FARID, H. 2004. A digital technique for art authentication. *Proc. Nat. Acad. Sci. 101*, 49, 17006–17010.

MA, W. AND MANJUNATH, B. 1997. Netra: A toolbox for navigating large image databases. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*.

MA, W.-Y. AND MANJUNATH, B. 1998. Texture thesaurus for browsing large aerial photographs. *J. Amer. Soc. Inf. Sci. 49*, 7, 633–648.

MAITRE, H., SCHMITT, F., AND LAHANIER, C. 2001. 15 years of image processing and the fine arts. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*.

MALIK, J., BELONGIE, S., LEUNG, T. K., AND SHI, J. 2001. Contour and texture analysis for image segmentation. *Int. J. Comput. Vision 43*, 1, 7–27.

MALIK, J. AND PERONA, P. 1990. Preattentive texture discrimination with early vision mechanisms. *J. Optical Soc. Amer. A 7*, 5, 923–932.

MALLOWS, C. L. 1972. A note on asymptotic joint normality. *Ann. Math. Statist. 43*, 2, 508–515.

MANJUNATH, B. AND MA, W.-Y. 1996. Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell. 18*, 8, 837–842.

MANJUNATH, B. S., OHM, J.-R., VASUDEVAN, V. V., AND YAMADA, A. 2001. Color and texture descriptors. *IEEE Trans. Circ. Syst. Video Technol. 11*, 6, 703–715.

MARSICOI, M. D., CINQUE, L., AND LEVIALDI, S. 1997. Indexing pictorial documents by their content: A survey of current techniques. *Image Vision Comput. 15*, 2, 119–141.

MARTINEZ, K., CUPITT, J., SAUNDERS, D., AND PILLAY, R. 2002. Ten years of art imaging research. *Proc. IEEE 90*, 28–41.

MATHIASSEN, J. R., SKAVHAUG, A., AND BO, K. 2002. Texture similarity measure using Kullback-Leibler divergence between gamma distributions. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

MCLACHLAN, G. AND PEEL, D. 2000. *Finite Mixture Models*. Wiley-Interscience.

MEHROTRA, R. AND GARY, J. E. 1995. Similar-Shape retrieval in shape data management. *IEEE Comput. 28*, 9, 57–62.

MELZER, T., KAMMERER, P., AND ZOLDA, E. 1998. Stroke detection of brush strokes in protrait miniatures using a semi-parametric and a model based approach. In *Proceedings of the International IEEE Conference on Pattern Recognition (ICPR)*.

MIKOLAJCZK, K. AND SCHMID, C. 2003. A performance evaluation of local descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

MILLER, G. 1995. Wordnet: A lexical database for English. *Comm. ACM 38*, 11, 39–41.

MIRSKY, S.. 2006. Computers get the picture. *Sci. Amer.* (Nov.).

MITRA, P., MURTHY, C., AND PAL, S. 2002. Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 3, 301–312.

MOKHTARIAN, F. 1995. Silhouette-Based isolated object recognition through curvature scale space. *IEEE Trans. Pattern Anal. Mach. Intell. 17*, 5, 539–544.

MONAY, F. AND GATICA-PEREZ, D. 2003. On image auto-annotation with latent space models. In *Proceedings of the ACM International Conference on Multimedia*.

MORI, G. AND MALIK, J. 2003. Recognizing objects in adversarial clustter: Breaking a visual captcha. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

MUKHERJEA, S., HIRATA, K., AND HARA, Y. 1999. Amore: A World Wide Web image retrieval engine. In *Proceedings of the International World Wide Web Conference (WWW)*.

MULLER, H., MARCHAND-MAILLET, S., AND PUN, T. 2002. The truth about Corel—Evaluation in image retrieval. In *Proceedings of the International Conference on Video Retrieval (CIVR)*. Lecture Notes in Computer Science, vol. 2383. Springer, 36–45.

MULLER, H., MICHOUX, N., BANDON, D., AND GEISSBUHLER, A. 2004. A review of content-based image retrieval systems in medical applications—Clinical benefits and future directions. *Int. J. Medical Inf. 73*, 1, 1–23.

MULLER, H., MULLER, W., SQUIRE, D. M., MARCHAND-MAILLET, S., AND PUN, T. 2001. Performance evaluation in content-based image retrieval: Overview and proposals. *Pattern Recogn. Lett. 22*, 5, 593–601.

MULLER, H., PUN, T., AND SQUIRE, D. 2004. Learning from user behavior in image retrieval: Application of market basket analysis. *Int. J. Comput. Vision 56*, 1–2, 65–77.

NAGAMINE, T., JAIMES, A., OMURA, K., AND HIRATA, K. 2004. A visuospatial memory cue system for meeting video retrieval. In *Proceedings of the ACM International Conference on Multimedia* (demonstration).

NAKANO, K. AND TAKAMICHI, E. 2003. An image retrieval system using FPGAs. In *Proceedings of the Asia South Pacific Design Automation Conference*.

NAKAZATO, M., DAGLI, C., AND HUANG, T. 2003. Evaluating group-based relevance feedback for content-based image retrieval. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*.

NATSEV, A., NAPHADE, M. R., AND TESIC, J. 2005. Learning the semantics of multimedia queries and concepts from a small number of examples. In *Proceedings of the ACM International Conference on Multimedia*.

NATSEV, A., RASTOGI, R., AND SHIM, K. 2004. Walrus: A similarity retrieval algorithm for image databases. *IEEE Trans. Knowl. Data Eng. 16*, 3, 301–316.

NATSEV, A. AND SMITH, J. 2002. A study of image retrieval by anchoring. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*.

NG, T.-T., CHANG, S.-F., HSU, J., XIE, L., AND TSUI, M.-P. 2005. Physics-Motivated features for distinguishing photographic images and computer graphics. In *Proceedings of the ACM International Conference on Multimedia*.

PAINTER, T. H., DOZIER, J., ROBERTS, D. A., DAVIS, R. E., AND GREEN, R. O. 2003. Retrieval of subpixel snow-covered area and grain size from imaging spectrometer data. *Remote Sens. Env. 85*, 1, 64–77.

PANDA, N. AND CHANG, E. Y. 2006. Efficient top-k hyperplane query processing for multimedia information retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

PENTLAND, A., PICARD, R., AND SCLAROFF, S. 1994. Photobook: Tools for content-based manipulation of image databases. In *Proceedings of the Conference on Storage and Retrieval for Image and Video Database II, SPIE*, San Jose, CA.

PETRAGLIA, G., SEBILLO, M., TUCCI, M., AND TORTORA, G. 2001. Virtual images for similarity retrieval in image databases. *IEEE Trans. Knowl. Data Eng. 13*, 6, 951–967.

PETRAKIS, E. AND FALOUTSOS, A. 1997. Similarity searching in medical image databases. *IEEE Trans. Knowl. Data Eng. 9*, 3, 435–447.

PETRAKIS, E. G. M., DIPLAROS, A., AND MILIOS, E. 2002. Matching and retrieval of distorted and occluded shapes using dynamic programming. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 4, 509–522.

PETRAKIS, E. G. M., FALOUTSOS, C., AND LIN, K. I. 2002. Imagemap: An image indexing method based on spatial similarity. *IEEE Trans. Knowl. Data Eng. 14*, 5, 979–987.

PHOTO.NET. 1993. Photonet homepage. http://www.photo.net.

PHOTO.NET(RATINGSYSTEM). Photonet standards. 1993. http://www.photo.net/gallery/photocritique/standards.

PI, M., MANDAL, M. K., AND BASU, A. 2005. Image retrieval based on histogram of fractal parameters. *IEEE Trans. Multimed. 7*, 4, 597–605.

PICASA. 2004. Picasa homepage. http://picasa.google.com/.

PORTILLA, J. AND SIMONCELLI, E. 2000. A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vision 40*, 1, 49–71.

QUACK, T., MONICH, U., THIELE, L., AND MANJUNATH, B. S. 2004. Cortina: A system for largescale, content-based Web image retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

RODDEN, K., BASALAJ, W., SINCLAIR, D., AND WOOD, K. 2001. Does organization by similarity assist image browsing? In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

RODDEN, K. AND WOOD, K. 2003. How do people manage their digital photographs? In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

ROWE, N. C. 2002. Marie-4: A high-recall, self-improving Web crawler that finds images using captions. *IEEE Intell. Syst. 17*, 4, 8–14.

RUBNER, Y., TOMASI, C., AND GUIBAS, L. J. 2000. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision 40*, 99–121.

RUBNER, Y., TOMASI, C., AND GUIBAS, L. J. 1999. A metric for distribution with applications to image databases. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

RUI, Y., HUANG, T., AND CHANG, S.-F. 1999. Image retrieval: Current techniques, promising directions and open issues. *J. Visual Commun. Image Represent. 10*, 1, 39–62.

RUI, Y., HUANG, T., AND MEHROTRA, S. 1997. Content-based image retrieval with relevance feedback in Mars. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*.

RUI, Y. AND HUANG, T. S. 2000. Optimizing learning in image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

RUI, Y., HUANG, T. S., ORTEGA, M., AND MEHROTRA, S. 1998. Relevance feedback: A power tool in interactive content-based image retrieval. *IEEE Trans. Circ. Syst. Video Technol. 8*, 5, 644–655.

SABLATNIG, R., KAMMERER, P., AND ZOLDA, E. 1998. Hierarchical classification of paintings using face- and brush stroke models. In *Proceedings of the International IEEE Conference on Pattern Recognition (ICPR)*.

SAUX, B. L. AND BOUJEMAA, N. 2002. Unsupervised robust clustering for image database categorization. In *Proceedings of the International IEEE Conference on Pattern Recognition (ICPR)*.

SCHAFFALITZKY, F. AND ZISSERMAN, A. 2001. Viewpoint invariant texture matching andwide baseline stereo. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

SCHMID, C. AND MOHR, R. 1997. Local grayvalue invariants for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell. 19*, 5, 530–535.

SCHRODER, M., REHRAUER, H., SEIDEL, K., AND DATCU, M. 2000. Interactive learning and probabilistic retrieval in remote sensing image archives. *IEEE Trans. Geosci. Remote Sens. 38*, 5, 2288–2298.

SEBE, N., LEW, M. S., AND HUIJSMANS, D. P. 2000. Toward improved ranking metrics. *IEEE Trans. Pattern Anal. Mach. Intell. 22*, 10, 1132–1141.

SEBE, N., LEW, M. S., ZHOU, X., HUANG, T. S., AND BAKKER, E. 2003. The state of the art in image and video retrieval. In *Proceedings of the International Conference on Video Retrieval (CIVR)*.

SHANABLEH, T. AND GHANBARI, M. 2000. Heterogeneous video transcoding to lower spatio-temporal resolutions and different encoding formats. *IEEE Trans. Multimed. 2*, 2, 101–110.

SHI, J. AND MALIK, J. 2000. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell. 22*, 8, 888–905.

SHIRAHATTI, N. V. AND BARNARD, K. 2005. Evaluating image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

SLASHDOT. 2005. Searching by image instead of keywords. *Slashdot News* (May).

SMEATON, A. F. AND OVER, P. 2003. Benchmarking the effectiveness of information retrieval tasks on digital video. In *Proceedings of the International Conference on Video Retrieval (CIVR)*.

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

SMITH, J. AND CHANG, S.-F. 1997a. Integrated spatial and feature image query. *IEEE Trans. Knowl. Data Eng. 9*, 3, 435–447.

SMITH, J. AND CHANG, S.-F. 1997b. Visualseek: A fully automated content-based image query system. In *Proceedings of the ACM International Conference on Multimedia*.

SMOLKA, B., SZCZEPANSKI, M., LUKAC, R., AND VENETSANOPOULOS, A. N. 2004. Robust color image retrieval for the World Wide Web. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

SNOEK, C. G. M. AND WORRING, M. 2005. Multimodal video indexing: A review of the state-of-the-art. *Multimed. Tools Appl. 25*, 1, 5–35.

STAEDTER, T. 2006. Digital pics 'read' by computer. http://dsc.discovery.com/news/2006/11/09/images_tec.html.

SU, Z., ZHANG, H.-J., LI, S., AND MA, S. 2003. Relevance feedback in content-based image retrieval: Bayesian framework, feature subspaces, and progressive learning. *IEEE Trans. Image Process. 12*, 8, 924–937.

SWAIN, M. AND BALLARD, B. 1991. Color indexing. *Int. J. Comput. Vision 7*, 1, 11–32.

SWETS, D. AND WENG, J. 1996. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell. 18*, 8, 831–836.

TERRAGALLERIA. 2001. Terragalleria homepage. http://www.terragalleria.com.

THEOHARATOS, C., LASKARIS, N. A., ECONOMOU, G., AND FOTOPOULOS, S. 2005. A generic scheme for color image retrieval based on the multivariate Wald-Wolfowitz test. *IEEE Trans. Knowl. Data Eng. 17*, 6, 808–819.

TIAN, Q., SEBE, N., LEW, M. S., LOUPIAS, E., AND HUANG, T. S. 2001. Image retrieval using wavelet-based salient points. *J. Electron. Imag. 10*, 4, 835–849.

TIEU, K. AND VIOLA, P. 2004. Boosting image retrieval. *Int. J. Comput. Vision 56*, 1–2, 17–36.

TISHBY, N., PEREIRA, F., AND BIALEK, W. 1999. The information botflencek method. In *Proceedings of the Allerton Conference on Communication and Computation*.

TONG, S. AND CHANG, E. 2001. Support vector machine active learning for image retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

TOPE, A. S. AND ENSER, P. G. P. 2000. Design and implementation factors in electronic image retrieval systems. Res. Rep. 105, Library and Information Commission.

TORRES, R. S., SILVA, C. G., MEDEIROS, C. B., AND ROCHA, H. V. 2003. Visual structures for image browsing. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*.

TRECVID. 2001. TRECVID homepage. http://www-nlpir.nist.gov/projects/trecvid.

TU, Z. AND ZHU, S.-C. 2002. Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 5, 657–673.

TUYTELAARS, T. AND VAN GOOL, L. 1999. Content-Based image retrieval based on local affinely invariant regions. In *Proceedings of the International Conference on Visual Information Systems (VISUAL)*.

UNSER, M. 1995. Texture classification and segmentation using wavelet frames. *IEEE Trans. Image Process. 4*, 11, 1549–1560.

VAILAYA, A., FIGUEIREDO, M. A. T., JAIN, A. K., AND ZHANG, H.-J. 2001. Image classification for content-based indexing. *IEEE Trans. Image Process. 10*, 1, 117–130.

VASCONCELOS, N. 2004. On the efficient evaluation of probabilistic similarity functions for image retrieval. *IEEE Trans. Inf. Theory 50*, 7, 1482–1496.

VASCONCELOS, N. AND LIPPMAN, A. 2000a. Learning from user feedback in image retrieval systems. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*.

VASCONCELOS, N. AND LIPPMAN, A. 2000b. A probabilistic architecture for content-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

VASCONCELOS, N. AND LIPPMAN, A. 2005. A multiresolution manifold distance for invariant image similarity. *IEEE Trans. Multimed. 7*, 1, 127–142.

VELIVELLI, A., NGO, C.-W., AND HUANG, T. S. 2004. Detection of documentary scene changes by audio-visual fusion. In *Proceedings of the International Conference on Video Retrieval (CIVR)*.

VETRO, A., CHRISTOPOULOS, C., AND SUN, H. 2003. Video transcoding architectures and techniques: An overview. *IEEE Signal Process. Mag. 20*, 2, 18–29.

VINAY, V., COX, I. J., MILIC-FRAYLING, N., AND WOOD, K. 2004. Evaluating relevance feedback and display strategies for searching on small displays. In *Proceedings of the 7th International Symposium on String Processing Information Retrieval*. Lecture Notes in Computer Science, vol. 3246, 131–133.

VINAY, V., COX, I. J., MILIC-FRAYLING, N., AND WOOD, K. 2005. Evaluating relevance feedback algorithms for searching on small displays. In *Proceedings of the European Conference on IR Research (ECIR)*. Lecture Notes in Computer Science, vol. 3408. Springer, 185–199.

VON AHN, L. AND DABBISH, L. 2004. Labeling images with a computer game. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

WANG, J., LI, J., AND WIEDERHOLD, G. 2001. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans. Pattern Anal. Mach. Intell. 23*, 9, 947–963.

WANG, J., WIEDERHOLD, G., FIRSCHEIN, O., AND WEI, S. 1998. Content-Based image indexing and searching using Daubechies' wavelets. *Int. J. Digital Librar. 1*, 4, 311–328.

WANG, J. Z., BOUJEMAA, N., DEL BIMBO, A., GEMAN, D., HAUPTMANN, A., AND TESIC, J. 2006. Diversity in multimedia information retrieval research. In *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR) at the International Conference on Multimedia*.

WANG, J. Z., LI, J., GRAY, R. M., AND WIEDERHOLD, G. 2001. Unsupervised multiresolution segmentation for images with low depth of field. *IEEE Trans. Pattern Anal. Mach. Intell. 23*, 1, 85–90.

WANG, X.-J., MA, W.-Y., HE, Q.-C., AND LI, X. 2004a. Grouping Web image search result. In *Proceedings of the ACM International Conference on Multimedia*.

WANG, X. J., MA, W. Y., XUE, G. R., AND LI, X. 2004b. Multi-Model similarity propagation and its application for Web image retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

WANG, Y. H. 2003. Image indexing and similarity retrieval based on spatial relationship model. *Inf. Sci. Inf. Comput. Sci. 154*, 1-2, 39–58.

WANG, Z., CHI, Z., AND FENG, D. 2002. Fuzzy integral for leaf image retrieval. In *Proceedings of the IEEE International Conference on Fuzzy Systems*.

WEBE, M., WELLING, M., AND PERONA, P. 2000. Unsupervised learning of models for recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

WESTON, J., MUKHERJEE, S., CHAPELLE, O., PONTIL, M., POGGIO, T., AND VAPNIK, V. 2000. Feature selection for SVMS. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*.

WILSON, R. AND HANCOCK, E. 1997. Structural matching by discrete relaxation. *IEEE Trans. Pattern Anal. Mach. Intell. 19*, 6, 634–648.

WOLFSON, H. AND RIGOUTSOS, I. 1997. Geometric hashing: An overview. *IEEE Trans. Comput. Sci. Eng. 4*, 4, 10–21.

WOODROW, E. AND HEINZELMAN, W. 2002. Spin-It: A data centric routing protocol for image retrieval in wireless networks. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*.

WU, G., CHANG, E. Y., AND PANDA, N. 2005. Formulating context-dependent similarity functions. In *Proceedings of the ACM International Conference on Multimedia*.

WU, H., LU, H., AND MA, S. 2004. Willhunter: Interactive image retrieval with multilevel relevance measurement. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*.

WU, P. AND MANJUNATH, B. S. 2001. Adaptive nearest neighbor search for relevance feedback in large image databases. In *Proceedings of the ACM International Conference on Multimedia*.

WU, Y., CHANG, E. Y., CHANG, K. C. C., AND SMITH, J. R. 2004. Optimal multimodal fusion for multimedia data analysis. In *Proceedings of the ACM International Conference on Multimedia*.

WU, Y., TIAN, Q., AND HUANG, T. S. 2000. Discriminant-EM algorithm with application to image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

XIE, X., LIU, H., GOUMAZ, S., AND MA, W.-Y. 2005. Learning user interest for image browsing on small-form-factor devices. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

XING, E., NG, A., JORDAN, M., AND RUSSELL, S. 2003. Distance metric learning, with application to clustering with side-information. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*.

YANG, C., DONG, M., AND FOTOUHI, F. 2005a. Region based image annotation through multiple-instance learning. In *Proceedings of the ACM International Conference on Multimedia*.

YANG, C., DONG, M., AND FOTOUHI, F. 2005b. Semantic feedback for interactive image retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

YEE, K.-P., SWEARINGEN, K., LI, K., AND HEARST, M. 2003. Faceted metadata for image search and browsing. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

YU, J., AMORES, J., SEBE, N., AND TIAN, Q. 2006. Toward robust distance metric analysis for similarity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

YU, S. X. AND SHI, J. 2004. Segmentation given partial grouping constraints. *IEEE Trans. Pattern Anal. Mach. Intell. 26*, 2, 173–183.

ZHAI, Y., YILMAZ, A., AND SHAH, M. 2005. Story segmentation in news videos using visual and textual cues. In *Proceedings of the ACM International Conference on Multimedia*.

ZHANG, D.-Q. AND CHANG, S.-F. 2004. Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In *Proceedings of the ACM International Conference on Multimedia*.

ZHANG, H., RAHMANI, R., CHOLLETI, S. R., AND GOLDMAN, S. A. 2006. Local image representations using pruned salient points with applications to CBIR. In *Proceedings of the ACM International Conference on Multimedia*.

ZHANG, H. J., WENYIN, L., AND HU, C. 2000. IFIND—A system for semantics and feature based image retrieval over Internet. In *Proceedings of the ACM International Conference on Multimedia*.

ZHANG, L., CHEN, L., JING, F., DENG, K., AND MA, W. Y. 2006. Enjoyphoto—-A vertical image search engine for enjoying high-quality photos. In *Proceedings of the ACM International Conference on Multimedia*.

ZHANG, L., CHEN, L., LI, M., AND ZHANG, H.-J. 2003. Automated annotation of human faces in family albums. In *Proceedings of the ACM International Conference on Multimedia*.

ZHANG, Q., GOLDMAN, S. A., YU, W., AND FRITTS, J. E. 2002. Content-based image retrieval using multiple-instance learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

ZHANG, R. AND ZHANG, Z. 2004. Hidden semantic concept discovery in region based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

ZHANG, Y., BRADY, M., AND SMITH, S. 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Medical Imag. 20*, 1, 45–57.

ZHAO, W., CHELLAPPA, R., PHILLIPS, P. J., AND ROSENFELD, A. 2003. Face recognition: A literature survey. *ACM Comput. Surv. 35*, 4, 399–458.

ZHENG, B., MCCLEAN, D. C., AND LU, X. 2006. Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC Bioinf. 7*, 58.

ZHENG, X., CAI, D., HE, X., MA, W.-Y., AND LIN, X. 2004. Locality preserving clustering for image database. In *Proceedings of the ACM International Conference on Multimedia*.

ZHOU, D., WESTON, J., GRETTON, A., BOUSQUET, O., , AND SCHOLKOPF, B. 2003. Ranking on data manifolds. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*.

ZHOU, X. S. AND HUANG, T. S. 2001a. Comparing discriminating transformations and SVM for learning during multimedia retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

ZHOU, X. S. AND HUANG, T. S. 2001b. Small sample learning during multimedia retrieval using biasmap. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

ZHOU, X. S. AND HUANG, T. S. 2002. Unifying keywords and visual contents in image retrieval. *IEEE Multimed. 9*, 2, 23–33.

ZHOU, X. S. AND HUANG, T. S. 2003. Relevance feedback in image retrieval: A comprehensive review. *Multimed. Syst. 8*, 536–544.

ZHU, L., ZHANG, A., RAO, A., AND SRIHARI, R. 2000. Keyblock: An approach for content-based image retrieval. In *Proceedings of the ACM International Conference on Multimedia*.

ZHU, S.-C. AND YUILLE, A. 1996. Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell. 18*, 9, 884–900.

Received November 2006; revised June 2007; accepted July 2007

Image Retrieval: Ideas, Influences, and Trends of the New Age - Addendum

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang
The Pennsylvania State University

March 13, 2007

1 Publication Trends

We analyze recent publication trends in CBIR and annotation via two exercises, with Google Scholar as aid. The first of these is an analysis of which venues/journals have carried the most CBIR-related work and what the impact is, and which sub-topics generated the most publication count and impact in the last five years. The second one involves generating subtopic-wise time-series capturing trends in publication over the last eleven years.

We query on the phrase "image OR images OR picture OR pictures OR content-based OR indexing OR 'relevance feedback' OR annotation ", year 2000 onwards, for publications in the journals - IEEE T. Pattern Analysis and Machine Intelligence (PAMI), IEEE T. Image Processing (TIP), IEEE T. Circuits and Systems for Video Technology (CSVT), IEEE T. Multimedia (TOM), J. Machine Learning Research (JMLR), International J. Computer Vision (IJCV), Pattern Recognition Letters (PRL), and ACM Computing Surveys (SURV) and conferences - IEEE Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), IEEE International Conference on Image Processing (ICIP), ACM Multimedia (MM), ACM SIG Information Retrieval (IR), and ACM Human Factors in Computing Systems (CHI). Relevant papers among the top 100 results in each of these searches are used for the study. Google Scholar presents results roughly in decreasing order of citations (again, only rough approximations to the actual numbers). Limiting search to the top few papers translates to reporting statistics on work with noticeable impact. We gathered statistics on two parameters, (1) publishing venue/journal, and (2) sub-topics of interest. These trends are reported in terms of (a) number of papers, and (b) total number of citations. Plots of these scores are presented in Fig. 1 and Fig. 2. Note that the tabulation is not mutually exclusive (i.e. one paper can have contributions in multiple sub-topics such as 'Learning' and 'Region', and hence are counted under both headings), neither is it exhaustive or scientifically precise (Google's citation values may not be accurate). Nevertheless, these plots convey general trends in the relative impact

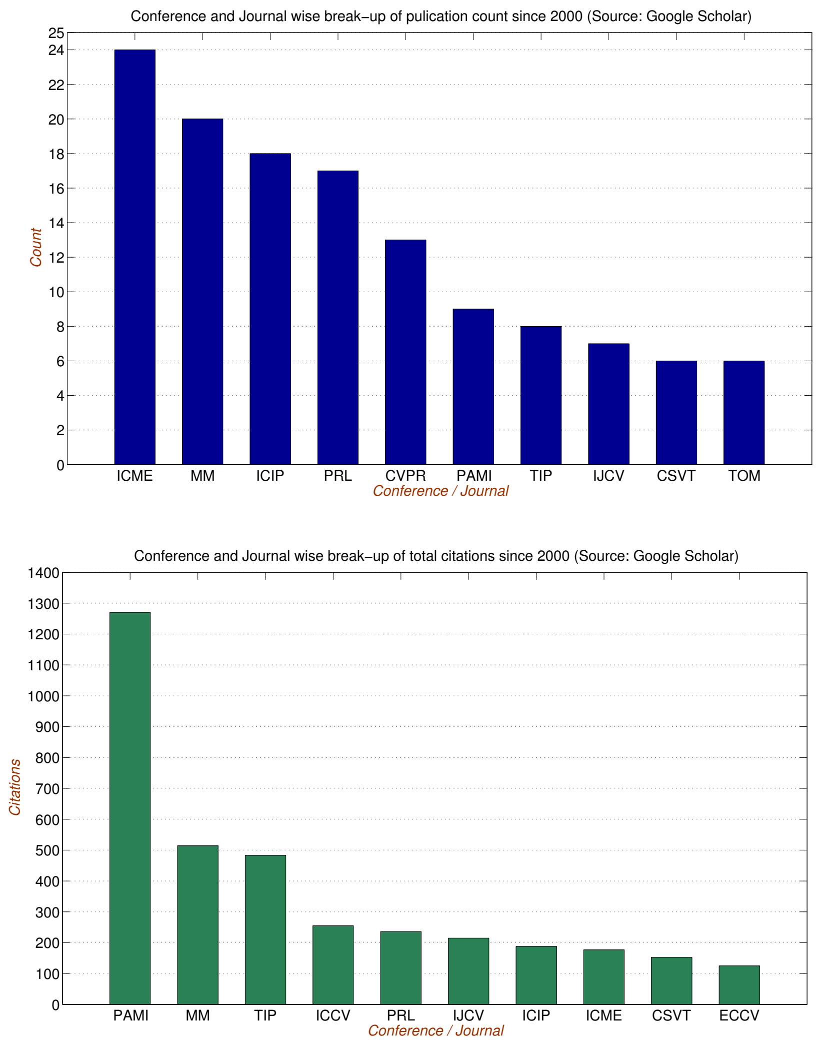
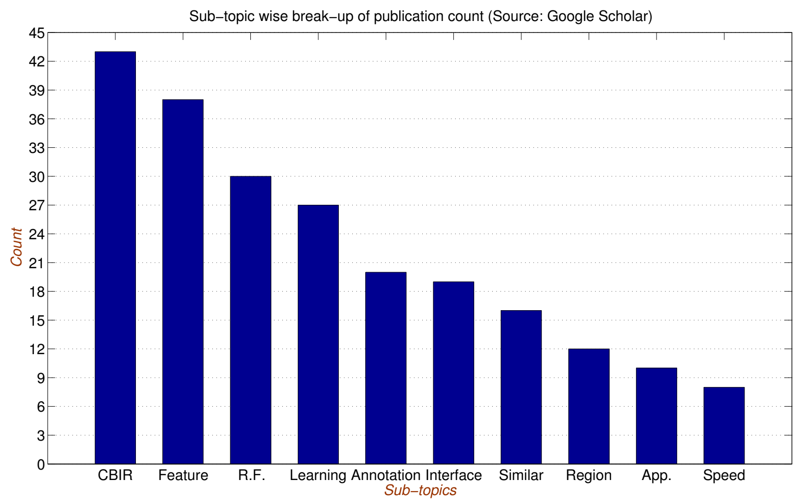


Figure 1: Conference-wise and journal-wise publication statistics on topics closely related to image retrieval, year 2000 onwards. *Top:* Publication counts. *Bottom:* Total citations.

of scholarly work. Readers are advised to use discretion when interpreting these results.

For the second experiment, we query Google Scholar for the phrase "image retrieval" for each year from 1995 to 2005, and note the publication count, say x. We then add a phrase corresponding to a CBIR-related technique, e.g., relevance feedback, and note the publication count again, say y. For each year



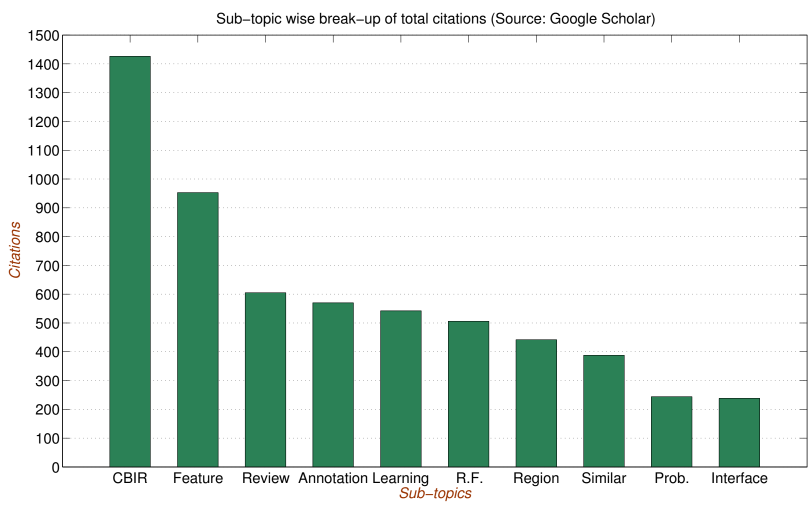


Figure 2: Publication statistics on sub-topics of image retrieval, 2000 onwards. *Top:* Publication Counts. *Bottom:* Total citations. *Abbreviations: Feature* - Feature Extraction, *R.F.* - Relevance Feedback, *Similar* - Image similarity measures, *Region* - Region based approaches, *App.* - Applications, *Prob.* - Probabilistic approaches, *Speed* - Speed and other performance enhancements.

and for each phrase, we take the ratio y/x representing the fraction of relevant publications. The time-series plot for eight such phrases, over the eleven years, can be seen in Fig. 3.

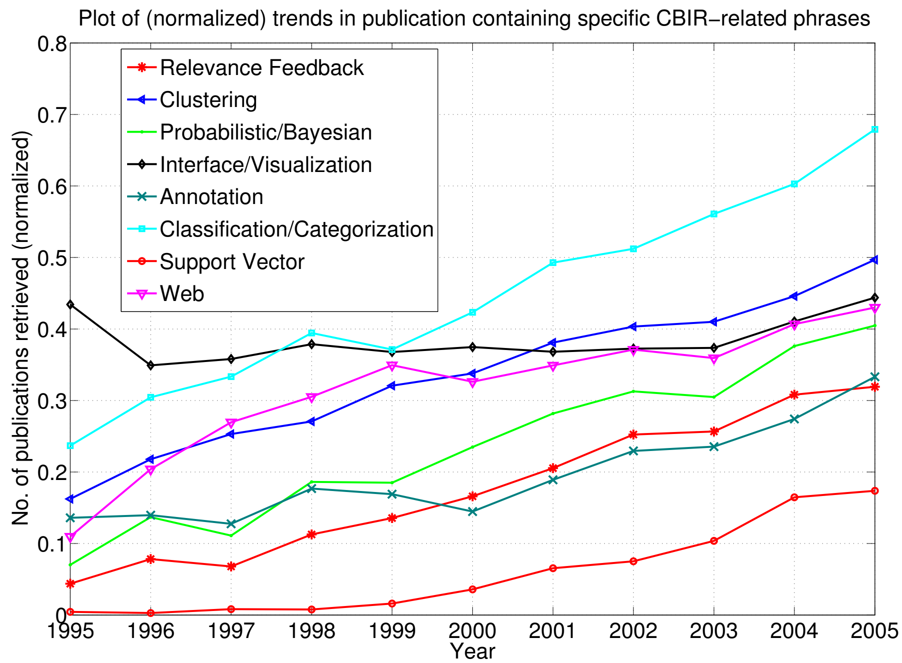


Figure 3: Normalized trends in publications containing "image retrieval" and corresponding phrases, as indexed by Google Scholar. Counts are normalized by the number of papers having "image retrieval" for the particular year.

2 Scientific Impact on Other Research Communities

The list of references in this paper is probably a good way to understand how diverse CBIR as a field is. There are at least 30 different well-known journals or proceedings where CBIR-related publications can be found, spanning at least eight different fields. In order to quantify this impact, we conduct a study. All the CBIR-related papers, cited in this work, are analyzed in the following manner. Let a set of CBIR-related fields be denoted as $\mathbf{F} = \{$*Multimedia (MM), Information Retrieval (IR), Digital Libraries/ World Wide Web (DL), Human-Computer Interaction (HCI), Language Processing (LN), Artificial Intelligence (including ML) (AI), Computer Vision (CV)*$\}$. Note the overlap among these fields, even though we treat them as distinct and non-overlapping for the sake of analysis. For each paper, we note what the core contribution is, including any new technique being introduced. For each such contribution, the core field it is associated with, $a \in \mathbf{F}$, is noted. For example, a paper that proposed a spectral clustering based technique for computing image similarity is counted under both

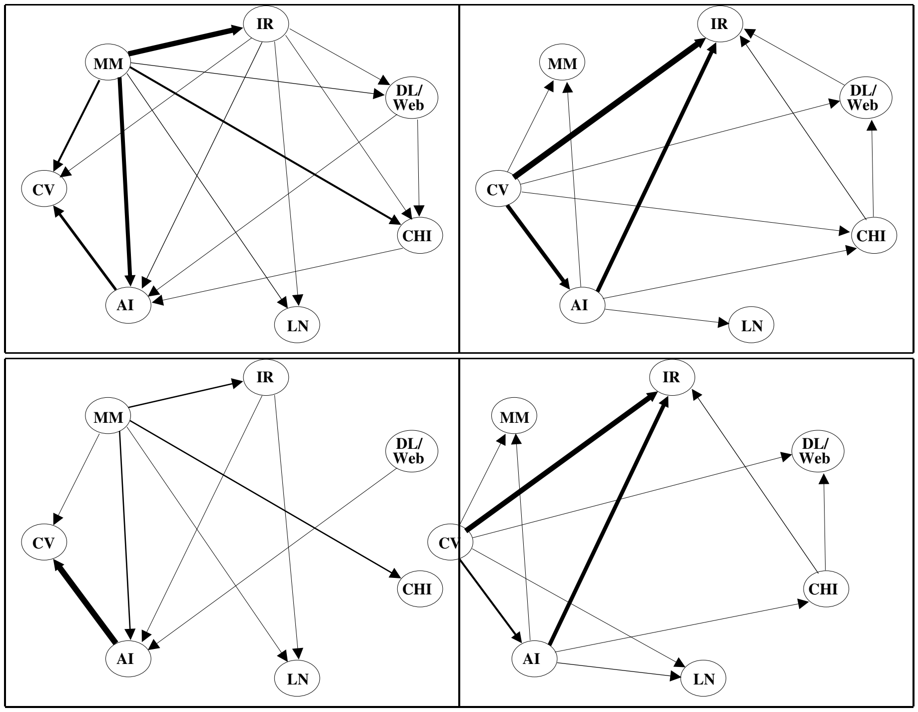


Figure 4: [Acronyms: MM := Multimedia, IR := Information Retrieval, DL := Digital Libraries/ World Wide Web, HCI := Human-Computer Interaction, LN := Language Processing, AI := Artificial Intelligence, and CV := Computer Vision]. Directed graphs representing inter-field impact induced by CBIR-related publications. An edge $a \rightarrow b$ implies publications at venue/journal concerning field b, having content concerning field a. We show oppositely directed edges between pairs of nodes, wherever significant, in the left and right graphs. *Top:* Edge thicknesses represent (relative) **publication count**. *Bottom:* Edge thicknesses represent (relative) **citations** as reported by Google Scholar.

CV and *AI*. Now, given the journal/venue where the paper was published, we note the field $b \in \mathbf{F}$ which it caters to, e.g., ACM SIGIR is counted under *IR* and ACM MIR Workshop is counted under both *IR* and *MM*. Over the 170 papers, we count the publication count and the Google Scholar citations for each $a \rightarrow b$ pair, $a \neq b$. The 7×7 matrices so formed ($|\mathbf{F}| = 7$) for count and citations are represented as directed graphs, as shown in Fig. 4. The thickness represents the publication or citation count, normalized by the maximum in the respective tables. Edges less than 5% of the maximum are not shown.

The basic idea behind constructing such graphs is to analyze how CBIR induces interests of one field of researchers in another field. A few trends are quite clear from the graphs. Most of the *MM*, *CV* and *AI* related work (i.e.

CBIR research whose content falls into these categories) has been published in *IR* venues and received high citations. At the same time, *AI* related work published in *CV* venues has generated considerable impact. We view this as a side-effect of CBIR research resulting in marriage of fields, communities, and ideas. But then again, there is little evidence of any mutual influence or benefits between the CV and CHI communities brought about by CBIR research.