

# Multimedia Information Retrieval at a Crossroad: Review and Outlook

Qing Li <sup>1</sup>      Yi Zhuang <sup>3</sup>      Jun Yang <sup>2</sup>      Yueting Zhuang <sup>3</sup>

<sup>1</sup> Department of Computer Engineering and Information Technology  
City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, HKSAR, CHINA

itqli@cityu.edu.hk

Tel: +852-2788-9695      Fax: +852-2788-8292

<sup>2</sup> School of Computer Science

Carnegie Mellon University, Pittsburgh, PA 15213, USA

juny@cs.cmu.edu

Tel: +1-412-268-2029      Fax: +1-412-268-6298

<sup>3</sup> Department of Computer Science

Zhejiang University, Hangzhou, 310027, CHINA

{zhuangyi, yzhuang}@cs.zju.edu.cn

Tel: +86-571-87951903      Fax: +86-571-87951947

# Multimedia Information Retrieval at a Crossroad: Review and Outlook

## **A B S T R A C T**

The proliferation of multimedia information calls for the research and systems on effective and efficient access to user-desired multimedia data. This chapter reviews three major techniques for multimedia information retrieval, namely text-based, content-based, and hybrid retrieval approaches, with their strengths and weaknesses compared. Moreover, as a critical component of large-scale multimedia information retrieval, works on high-performance indexing which can significantly improve retrieval efficiency are also reviewed. A brief introduction on the typical applications of multimedia retrieval is given, and the challenges are presented. Finally, the future research and applications of this area, particularly the recent trend on multi-modality retrieval systems, are discussed.

## **I N T R O D U C T I O N**

From late 1990s to early 2000s, the availability of powerful computing capability, large storage devices, high-speed networking, and especially the advent of Internet, led to a phenomenal growth of digital multimedia content in terms of size, diversity, and impact. As suggested by its name,

"multi-media" is a name given to a collection of data of multiple types, which include not only "traditional multimedia" such as images and videos, but also emerging media such as 3D graphics (like VRML objects) and Web animations (like Flash animations). Furthermore, relevant techniques have been developed for a growing number of applications, ranging from document editing software to digital libraries and many Web applications. For example, most people who ever used Microsoft Word have tried to insert pictures and diagrams into their documents, and they have the experience of watching online video clips such as movie trailers from websites such as YouTube.com. Multimedia data have been available in every corner of the digital world. With the huge volume of multimedia data, finding and accessing the **multimedia documents** that satisfy people's needs in an accurate and efficient manner becomes a non-trivial problem. This problem is referred to as *multimedia information retrieval*. The core of multimedia information retrieval is to compute the degree of relevance between users' information needs and multimedia data. A user's information need is expressed as a *query*, which can be in various forms such as a line of free text like "*Find me the photos of George Washington*", a few keywords like "*George Washington photo*", a media object like a sample picture of George Washington, or their combinations. On the other hand, multimedia data are represented using a certain form of summarization, typically called **index**, which is directly matched against

queries. Similar to a query, the index can take a variety of forms, including keywords, visual features such as color histogram and motion vector, depending on the data and task characteristics.

For textual documents, mature **information retrieval (IR)** technologies have been developed and successfully applied in commercial systems such as Web search engines. In comparison, the research on multimedia retrieval is still in its early stage. Unlike textual data, which can be well represented by term vectors that are descriptive of data semantics, multimedia data lack an effective, semantic-level representation that can be computed automatically, which makes multimedia retrieval a much harder research problem. On the other hand, the diversity and complexity of multimedia data offer new opportunities for the retrieval task to be leveraged by the techniques in other research areas. In fact, research on multimedia retrieval has been initiated and investigated by researchers from areas of **multimedia database**, computer vision, natural language processing, human-computer interaction, etc. Overall, it is currently a very active research area that has many interactions with other areas.

In the coming sections, we will overview the techniques for multimedia information retrieval, followed by a review on the applications and challenges in this area. Then, the future trends will be discussed, and some important terms in this area are defined at the end of this chapter.

## **M U L T I M E D I A   R E T R I E V A L   T E C H N I Q U E S**

Despite the various techniques proposed in literature, there exist three major approaches to multimedia retrieval, namely text-based approach, content-based approach, and hybrid approach. Their main difference lies in the type of index used for retrieval: the first approach uses text (keywords) as index, the second one uses low-level features extracted from multimedia data, and the last one uses the combination of text and low-level features. As a result, they differ from each other in many other aspects ranging from feature extraction to similarity measures.

### **Text-based Multimedia Retrieval**

Text-based multimedia retrieval approaches apply mature information retrieval (IR) techniques to the domain of multimedia retrieval. A typical text-IR method matches text queries issued by users with descriptive keywords extracted from documents. To use this method for multimedia retrieval, textual descriptions in the form of "bag of keywords" need to be extracted to describe multimedia objects, and user queries must be expressed as a set of keywords. Given the text descriptions and text queries, multimedia retrieval boils down to a text-IR problem. In early years such descriptions were usually obtained by manually annotating the multimedia data with keywords (Tamura and Yokoya, 1984). This approach is not scalable to large data if the number of human annotators is limited,

but is applicable if the annotation task is shared among a large population of users. This is the case of several image/video sharing websites, such as YouTube.com and Flickr.com, where users add (keyword) tags on their photos or videos such that they can be found by keyword search. The vulnerability to human bias is always an issue with manual annotations. There have been also proposals from computer vision and pattern recognition areas on automatically annotating the images and videos with keywords based on their low-level visual/audio features (Barnard et al. 2003, Jeon et al. 2004). Most of these approaches involve supervised or unsupervised machine learning, which tries to map low-level features into descriptive keywords. However, due to the large gap between multimedia data form (e.g., pixels, digits) and their semantic meanings, these approaches cannot produce high-quality keyword annotations. Some of the systems are semi-automatic, attempting to propagate keywords from a set of initially annotated objects to other objects. In some other applications, descriptive keywords can be easily accessible for multimedia data. Particularly, for images and videos embedded in web-pages, the text surrounding them as well as the title of the web-pages usually provide good descriptions, an approach explored both in research (e.g., Smith and Chang (1997)) and also in commercial image/video search engines (e.g., Google Image Search).

Since relatively speaking keyword annotations can precisely capture the semantic meanings of multimedia data, text-based retrieval approach is effective in terms of retrieving multimedia data that are *semantically relevant* to the users' needs. Moreover, because many people find it convenient and effective to use text (keywords) to express their information requests, as demonstrated by the fact that most commercial search engines (e.g., Google) support text queries, this approach has the advantage of being amenable to average users. But the bottleneck of this approach is still on the acquisition of keyword annotations, especially when there is a large amount of data and a small number of users, since no techniques provide both efficiency and accuracy in acquiring annotations when they are not available.

### **Content-based Multimedia Retrieval**

The idea of **content-based retrieval** first came from the area of content-based image retrieval (CBIR) (Flickner et al. 1995; Smeulders et al. 2000). Gradually the idea has been applied to the retrieval tasks for other media types, resulting in content-based video retrieval (Hauptmann et al. 2002; Somliar, 1994) and content-based audio retrieval (Foote 1999). The word "content" here refers to the low-level representation of the data, such as pixels for Bitmap images, MPEG bit-streams for MPEG-format video, etc. Content-based retrieval, as opposed to text-based retrieval, exploits the

features that are (automatically) extracted from the low-level representation of the data, usually denoted as low-level features since they do not directly capture the high-level meanings of the data. (In a sense, text-based retrieval of documents is also "content-based", since keywords are extracted from the content of documents.) Obviously, the low-level features used for retrieval depend on the type of data to be retrieved. For example, color histogram is a typical feature for image retrieval, and motion vector is used for video retrieval, etc. Despite the heterogeneity of the features, in most cases they can be transformed into feature vector(s) which are typically high-dimensional and real-valued. Thus, the similarity between media objects can be measured by the distance of their respective feature vectors in the vector space under certain distance metrics. Various distance measures, such as Euclidean distance, histogram intersection, can be used as the similarity metrics. This has a correspondence to the vector-based model for (text) information retrieval, where a bag of keywords is also represented as a vector. The original feature space can be transformed into a manifold space in which the distance metric better captures the intrinsic similarity/dissimilarity between media objects, an approach seen in recent works such as (He et al. 2004).

Content-based retrieval also influences the way a query is composed. Since a media object is represented by its low-level feature vector(s), a query must be also transformed into a feature vector in order to match against the object. This results in **query-by-example** (QBE) (Flickner et al. 1995), a search paradigm where media objects such images or video clips are used as query examples to find other objects similar to them, where "similar " is defined mainly at perceptual levels (i.e., looks like, or sounds like). In this case, feature vector(s) extracted from the example object(s) are matched with the feature vectors of the media objects to be retrieved. The majority of content-based retrieval systems use QBE as its search paradigm. However, there are also content-based systems that use alternative ways to let users specify their intended low-level features, such as by selecting from some templates or a small set of feature options (i.e., "red", "black", or "blue"). See the PIPA Project (2005) as an example.

The features and similarity metrics used by many content-based retrieval systems are chosen heuristically and are therefore ad-hoc and unjustified. It is questionable that the features and metrics are optimal or close to optimal. Thus, there have been efforts seeking for theoretically justified retrieval approaches whose optimality is guaranteed under certain circumstances (Sebe et al., 2000). Cooper et al. (2005) suggest measuring image similarity using time and pictorial content. Many of these

approaches treat retrieval as a machine learning problem of finding the most effective (weighted) combination of features and similarity metrics to solve a particular query or a set of queries. Such learning can be done online in the middle of the retrieval process, based on users' feedback evaluations or automatically derived "pseudo" feedbacks. In fact, relevance feedback (Rui et al. 1998) has been one of the hot topics in content-based image retrieval. Recently, Kelly et al. (2003) proposed a concept of implicit relevance feedback (IRF) which is different from traditional relevance feedback methods that require users to explicitly give feedbacks. An extension of IRF was also studied in (Ryen et al. 2005). Off-line learning has also been used to find effective features/weights based on previous retrieval experiences. However, machine learning is unlikely to be the magic answer for content-based retrieval problem, due to the fact that it is impossible to have training data for basically an infinite number of queries and users are usually unwilling to give feedbacks.

Overall, content-based retrieval has the advantage of being fully automatic from the feature extraction to similarity computation, and thus scalable to real systems. With the query-by-example (QBE) search paradigm, it is also able to capture the perceptual aspects of multimedia data that cannot be easily depicted by text. The downside of content-based retrieval is mainly due to the so called "semantic gap" between low-

level features and the semantic meanings of the data. Given the fact that users prefer semantically relevant results, content-based methods suffer from the low precision/recall problem, which prevents them from being used in commercial systems. Another problem lies in the difficulty of finding a suitable example object to form an effective query, if the QBE paradigm is used.

### **Hybrid Multimedia Retrieval**

The hybrid multimedia retrieval is proposed partially because both text-based and content-based retrieval have their own limitations and they can be complementary to each other. Different from these two retrieval methodologies, hybrid multimedia retrieval is emerging as a new multimedia retrieval method which tries to combine the retrieval results returned by text-based method and by content-based method to enhance the retrieval effectiveness (Shen et al. 2006). In the image retrieval approach proposed by Zhou et al. (2007), for a given query, text-based retrieval is first conducted to get a set of candidate images, and then a content-based refinement process is performed based on the low-level features of candidate images. This hybrid approach has almost become the standard in video retrieval, since video data by nature contain text (from speech recognition, closed captions, etc), video frames (as images), and audio information. For example, Rong et al. (2006) proposed a

probabilistic framework for combining the results of multiple “retrieval experts”, including text retrieval, key-frame visual similarity, and so-called semantic concepts, to acquire a list of relevant video shots for a given query. Empirical results have shown that hybrid multimedia retrieval can be regarded as a promising retrieval method compared with the above two methods.

## **H I G H - P E R F O R M A N C E I N D E X**

In the early multimedia retrieval systems, the multimedia objects such as images or video were frequently stored as simple files in a directory or entries in a relational table. From a perspective of computational efficiency, both options exhibited poor performance because most file systems use sequential search within directories. Thus, as the size of the multimedia databases or collections grew from hundreds, to thousands, to millions of variable sized objects, the computers could not respond in an acceptable time period.

As feature vector extracted from media objects is multi- or high-dimensional, the indexing of multimedia data belongs to the high-dimensional index issue. There is a long stream of research for addressing the high-dimensional indexing problems [Böhm et al. 1998]. Existing techniques can be divided into four main categories.

The first category is based on data and space partitioning, hierarchical tree index structure (e.g., the R-tree [Guttman. 1984] and its variants [Beckmann et al. 1990]), etc. Although these methods generally perform well at low dimensionality, their performance deteriorates rapidly as the dimensionality increases and the degradation can be so bad that sequential scanning becomes more efficient due to the "dimensionality curse".

The second category is to represent original feature vectors using smaller, approximate representations (e.g., VA-file [Weber et al. 1998] and IQ-tree [Berchtold et al. 2000]), etc. The VA-file [Weber et al. 1998] accelerates the sequential scan by using data compression. Although the VA-file reduces the number of disk accesses, it incurs higher computational cost to decode the bit-strings, compute all the lower and some upper bounds on the distance to the query point, and determine the actual distances of candidate points. The IQ-tree [Berchtold et al. 2000] is also an indexing structure along the lines of the VA-file, which maintains a flat directory containing the minimum bounding rectangles of the approximate data representations.

The third category is to use a metric-based method [Chávez et al. 2001] as an alternative direction for high-dimensional indexing. Examples include MVP-Tree [Bozkaya T. 1997] and M-Tree [Ciaccia et al. 1997], etc.

The final category is the transformation-based high-dimensional indexing schemes, such as the Pyramid Technique [Berchtold et al. 1998]. The Pyramid Technique is efficient for window queries, but performs less satisfactorily for k-NN queries. Most recently, iDistance [Jagadish et al. 2005] are proposed to support B+-tree- based k-NN search. It is proposed by selecting some reference points in order to further prune the search region so as to improve the query efficiency. However the query efficiency of iDistance relies largely on clustering and partitioning the data and is significantly affected if the choice of partition scheme and reference data points is not appropriate.

## **A P P L I C A T I O N S   A N D   C H A L L E N G E S**

Though far from mature, multimedia retrieval techniques have been widely used in a number of applications. The most visible application is various Web search engines for images and video, such as Image Search and Video Search at Google.com, Blinx.com, as well as image and video sharing sites, such as YouTube.com and Flickr.com. The search facilities provided by these systems are text-based, implying that a text query is a better vehicle of users' information need than an example-based query. Content-based retrieval is not applicable here due to its low accuracy problem, which gets even worse due to the huge data volume. Web search engines acquire textual annotations of images or video automatically by analyzing the text

in Web pages, but the results for some popular queries may be manually crafted. In image and video sharing sites, such text annotations are provided in the form of tags collectively by a huge population of users. Because of the huge data volume on the Web, the relevant data to a given query can be enormous. Therefore, the search engines need to deal with the problem of "authoritativeness", namely determining how authoritative a piece of data is, besides the problem of relevance. In addition to the Web, there are many offline digital libraries, such as Microsoft Encarta Encyclopedia, that have the facilities for searching multimedia objects like images and video clips by text. The search is usually realized by matching manual annotations with text queries.

Multimedia retrieval techniques have also been applied to some narrow domains, such as news videos, sports videos, medical imaging. NIST TREC Video Retrieval Evaluation (TRECVID) has attracted many research efforts devoting to various retrieval tasks on broadcast news video based on automatic analysis of video content. Sports videos like basketball programs and baseball programs have been studied to support intelligent access and summarization (Zhang and Chang 2002). In the medical imaging area, for example, Liu et al. (2002) applied retrieval techniques to detect brain tumor from CT/MR images. Content-based techniques have achieved some level of success in these domains, because the data size is relatively small and domain-specific features can be crafted to capture the

idiosyncrasy of the data. Generally speaking, however, there is no killer application where content-based retrieval techniques can achieve a fundamental breakthrough.

The emerging applications of multimedia also raise new challenges for multimedia retrieval technologies. One of such challenges comes from the new media formats emerged in recent years, such as Flash animation, PowerPoint file, SMIL (Synchronized Multimedia Integration Language). These new formats demand specific retrieval methods for them. Moreover, their intrinsic complexity (some of them can recursively contain media components) brings up new research problems not addressed by current techniques. There have already been recent efforts devoted to these new medias, such as Flash animation retrieval (Yang et al. 2002a), and PowerPoint presentation retrieval. Another challenge rises from the idea of retrieving multiple types of media data in a uniform framework, which will be discussed in the next section.

## **F U T U R E   T R E N D S**

In a sense, most existing multimedia retrieval methods are not genuinely for "multi-media", but for a specific type (or modality) of non-textual data. There is, however, the need to design a real "multi-media" retrieval system that can handle multiple data modalities in a cooperative framework. First, in multimedia databases like the Web, different types of media objects co-

exist as an organic whole to convey the intended information. Naturally, users would be interested in seeing the complete information by accessing all the relevant media objects regardless of their modality, preferably, from a single query. For example, a user interested in a new car model would like to see the pictures of the car and meanwhile read articles on it. Sometimes, depending on the physical conditions such as networks and displaying devices, users may want to see a particular presentation of the information in appropriate modality(-ies). Furthermore, some data types such as video intrinsically consist of data of multiple modalities (audio, closed-caption, video images). It is advantageous to explore all these modalities and let them complement each other in order to obtain better retrieval effect. To sum up, a retrieval system that goes across different media types and integrates **multi-modality** information is highly desirable. Informedia (Hauptmann et al. 2002) is a well-known video retrieval system that successfully combines multi-modal features. Its retrieval function not only relies on the transcript generated from a speech recognizer and/or detected from overlaid text on screen, but also utilizes features such as face detection and recognition results, image similarity, etc. Statistical learning methods are widely used in Informedia to intelligently combine the various types of information. There are many other systems that integrate features from at least two modalities for retrieval purpose. For example, WebSEEK system (Smith and Chang 1997) extracts keywords from

the surrounding text of image and videos in web pages, which is used as their indexes in the retrieval process. Although the systems involve more than one media type, typically textual information plays the vital role in providing the (semantic) annotation of the other media types.

Systems featuring a higher degree of integration of multiple modalities are emerging. More recently, the MediaNet (Benitez et al. 2002) and multimedia thesaurus (MMT) (Tansley 1998) are proposed, both of which seek to provide a multimedia representation of semantic concept—a concept described by various media objects including text, image, video, etc—and establish the relationships among these concepts. MediaNet extends the notion of relationships to include even perceptual relationships among media objects.

In (Yang et al. 2002b), a very comprehensive and flexible model named *Octopus* is proposed to perform "aggressive" search of multi-modality data. It is based on a multi-faceted knowledge base represented by a layered graph model, which captures the relevance between media objects of any type from various perspectives, such as the similarity on low-level features, structural relationships such as hyperlinks, and semantic relevance. Link analysis techniques can be used to find the most relevant objects for any given object in the graph. This new model can accommodate knowledge from various sources, and it allows a query to be composed flexibly using either text or example objects or both.

Recently, in (Wu et al, 2005), a concept of Cross-Media retrieval is first proposed which tries to break the limitation of modalities of media objects. As an extension of multi-modality retrieval, a cross-media retrieval can be regarded as an unified multimedia retrieval paradigm by learning some latent semantic correlation between different types of media objects.

## **C O N C L U S I O N**

Multimedia information retrieval is a relatively new area that has been receiving more and more attention from various research areas like database, computer vision, natural language, machine learning, as well as from industry. Given the continuing growth of multimedia data, research in this area will expectedly become more active since it is critical to the success of various multimedia applications. However, technological breakthroughs and killer applications in this area are yet to come, and before that, multimedia retrieval techniques can hardly be migrated to commercial applications. The breakthrough in this area depends on the joint efforts from its related areas, and therefore it offers researchers opportunities to tackle the problem from different paths and with different methodologies.

## **R E F E R E N C E**

- Barnard, K., Duygulu, P., Freitas, N., Forsyth, D., Blei, D., and Jordan, M. (2003) Matching Words and Pictures. *Journal of Machine Learning Research*, Vol 3: 1107-1135.
- Beckmann, N., Kriegel, H.-P., Schneider, R., Seeger, B. (1990). The R\*-tree: An Efficient and Robust Access Method for Points and Rectangles, In *Proceedings of ACM SIGMOD Conference*, pp. 322-331.
- Benitez, A. B., Smith, J. R. and Chang, S. F. (2000) "MediaNet: A Multimedia Information Network for Knowledge Representation". *Proceeding of the SPIE 2000 Conference on Internet Multimedia Management Systems*, vol.4210.
- Berchtold, S., Bohm, C., and Kriegel, H.-P. (1998). The pyramid technique: Towards breaking the curse of dimensionality. In *Proceedings of SIGMOD Conference*. Pp. 142-153.
- Berchtold, S., Bohm, C., Kriegel, H.P., Sander, J., and Jagadish, H.V. (2000). Independent quantization: An index compression technique for high-dimensional data spaces. In *Proceedings of the 16th ICDE Conference*, pp. 577-588.
- Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceeding of the 7th International World Wide Web Conference*, pp. 107-117.

- Böhm, C., Berchtold S., Keim, D. (2001). Searching in High-dimensional Spaces: Index Structures for Improving the Performance of Multimedia Databases, *ACM Computing Surveys*, 33 (3).
- Bozkaya T. and Ozsoyoglu, M. (1997). Distance-based indexing for high-dimensional metric spaces. In *Proceedings of ACM SIGMOD Conference*, pages 357-368. 1997.
- Chávez, E., Navarro, G., Baeza-Yates, R., and J. Marroquín (2001), Searching in Metric Spaces, *ACM Computing Surveys*: 33(3), pp. 273-321.
- Ciaccia, P., Patella, M., and Zezula, P. (1997). M-trees: An efficient access method for similarity search in metric space. In *Proceedings of the 23rd VLDB Conference*, pages 426-435. 1997.
- 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 Computer*, 28(9): 23-32.
- Foote, J. (1999) An overview of audio information retrieval, *Multimedia Systems*, 7(1): 2-10.
- Guttman, A. (1984), R-tree: A dynamic index structure for spatial searching, In *Proceedings of the ACM SIGMOD Conference*, pp.47-54.

- Hauptmann, A., et al. (2002) Video Classification and Retrieval with the Informedia Digital Video Library System. Text Retrieval Conference (TREC02), Gaithersburg, MD.
- He, X-F., Ma, W-Y.,, and Zhang, H-J., (2004) Learning an Image Manifold for Retrieval. Proceedings of ACM conference on Multimedia 2004, pp. 10-16.
- Kelly, D. and Teevan, J. (2003). Implicit feedback for inferring user preference. SIGIR Forum. 37(2): 18-28.
- Jeon, J., Lavrenko, V., and Manmatha, R., (2003) Automatic Image Annotation and Retrieval using Cross-media Relevance Models, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 119-126.
- Jagadish, H.V., Ooi, B.C., Tan, K.L., Yu, C., Zhang, R.(2005). iDistance: An Adaptive B+-tree Based Indexing Method for Nearest Neighbor Search., ACM Transactions on Data Base Systems. 30(2), pp. 364-397.
- Liu, Y., Lazar, N. and Rothfus, W. (2002) Semantic-based Biomedical Image Indexing and Retrieval. International Conference on Diagnostic Imaging and Analysis (ICDIA 2002).
- Lu, Y, Hu, C, Zhu, X, Zhang, H, Yang Q, (2000) A Unified Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems. Proceeding of ACM Multimedia Conference, pp 31-38.

- NIST TREC Video Retrieval Evaluation. <http://www-nlpir.nist.gov/projects/trecvid/>
- The PIPA Project.  
<http://spade.ddns.comp.nus.edu.sg/viper/demos.html>. 2005.
- Rong Yan, Alexander G. Hauptmann: Probabilistic latent query analysis for combining multiple retrieval sources. SIGIR 2006: 324-331
- Ryen W. White, Ian Ruthven, Joemon M. Jose (2005). A study of factors affecting the utility of implicit relevance feedback. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 35-42.
- Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S. (1998) Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval, *IEEE Trans on Circuits and Systems for Video Technology*, Special Issue on Segmentation, Description, and Retrieval of Video Content, Vol 8, pp.644-655.
- Shen, H-T, Zhou, X-F and Bin, C., (2006) "Indexing and Integrating Multiple Features for WWW images". World Wide Web Journal, 9(3):343-364.
- Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. (2000) Content-based image retrieval at the end of the early years. IEEE

Transactions on Pattern Analysis and Machine Intelligence,  
22(12):1349-1380.

- SMIL (Synchronized Multimedia Integration Language)  
<http://www.w3.org/AudioVideo/>
- Smith JR, Chang SF (1997) Visually Searching the Web for Content.  
IEEE Multimedia Magazine, 4 (3): 12-20.
- Somliar, S.W., Zhang, H. et al. (1994) Content-Based Video Indexing  
and Retrieval, IEEE MultiMedia, 1(2): 62-72.
- Tamura, H., Yokoya, N. (1984) Image database systems: A survey.  
Pattern Recognition, 17(1):29-43.
- Tansley, R., "The Multimedia Thesaurus: An Aid for Multimedia  
Information Retrieval and Navigation", Master Thesis, Computer  
Science, University of Southampton, UK, 1998.
- The VIPER interactive image retrieval system.  
<http://www.bestpeer.com/viper/demos.html>, 2006
- Weber, R., Schek, H., and Blott, S. (1998). A quantitative analysis and  
performance study for similarity-search methods in high-  
dimensional spaces. In Proceedings of the 24th VLDB Conference, pp.  
194-205.
- Wu, F., Zhang, H., Zhuang, Y-T.(2006),: Learning Semantic  
Correlations for Cross-Media Retrieval. ICIP:1465-1468

- Yang, J., Li, Q., Liu W., Zhuang, Y. (2002a) FLAME: A Generic Framework for Content-based Flash Retrieval. ACM MM'2002 Workshop on Multimedia Information Retrieval, Juan-les-Pins, France.
- Yang, J., Li, Q., Zhuang, Y. (2002b) Octopus: Aggressive Search of Multi-Modality Data Using Multifaceted Knowledge Base. Proc. of 11th International Conference on World Wide Web, pp.54-64, Hawaii, USA.
- Zhang, D., Chang, S.F (2002) Event detection in baseball video using superimposed caption recognition. Proceeding of ACM Multimedia Conference, pp. 315-318.
- Zhou, Z-H., Dai, H-B., (2007): Exploiting Image Contents in Web Search. Proceeding of International Joint Conference on Artificial Intelligence. Pp. 2922-2927

## **T E R M   D E F I N I T I O N S**

**Multimedia Database:** A database system that is dedicated to the storage, management, and access of one or more media types, such as text, image, video, sound, diagram, etc. For example, an image database such as Corel Image Gallery that stores a large number of pictures and allow users to browse them or search them by keywords can be regarded as a multimedia database. An electronic encyclopedia such as Microsoft Encarta

Encyclopedia, which consists of tens of thousands of multimedia documents with text descriptions, photos, video clips, animations, is another typical example of multimedia database.

**Multimedia Document:** A multimedia document is a natural extension of a conventional textual document in the multimedia area. It is defined as a digital document that is composed of one or multiple media elements of different types (text, image, video, etc) as a logically coherent unit. A multimedia document can be a single picture or a single MPEG video file, but more often it is a complicated document such as a web page consisting of both text and images.

**Multimedia Information Retrieval (system):** Storage, indexing, search, and delivery of multimedia data such as images, videos, sounds, 3D graphics, or their combination. By the definition, it includes works on, for example, extracting descriptive features from images, reducing high-dimensional indexes into low-dimensional ones, defining new similarity metrics, efficient delivery of the retrieved data, etc. Systems that provide all or part of the above functionalities are multimedia retrieval systems. The Google image search engine is a typical example of such a system. A Video-on-demand site that allows people to search movies by their titles is another example.

**Information Retrieval (IR):** The research area that deals with the storage, indexing, organization of, search, and access to information items, typically textual documents. Although its definition includes multimedia retrieval (since information items can be multimedia), the conventional IR refers to the work on textual documents, including retrieval, classification, clustering, filtering, visualization, summarization, etc. The research on IR started nearly half century ago and it grew fast in the past 20 years with the efforts of librarians, information experts, researchers on artificial intelligence and other areas. A system for the retrieval of textual data is an IR system, such as all the commercial Web search engines.

**Index:** In the area of information retrieval, "index" is the representation or summarization of a data item that is used for matching with queries to obtain the similarity between the data and the query, or matching with the indexes of other data items. For example, keywords are frequently used indexes of textual documents, and color histogram is a common index of images. Indexes can be manually assigned or automatically extracted. The text description of an image is usually manually given, but its color histogram can be computed by programs.

**High-Dimensional index:** For content-based multimedia retrieval, the low-level features extracted from the media objects such as image, audio, etc is usually multi- or high-dimensional. The high-dimensional index is a scheme which can efficiently and effectively organize and order the features from a great number of the multimedia objects. The aim of it is to improve the performance of similarity search over large multimedia databases by significantly reducing the search region.

**Content-based Retrieval:** An important retrieval method for multimedia data, which use the low-level features (automatically) extracted from the data as the indexes to match with queries. Content-based image retrieval is a good example. The specific low-level features used depend on the data type: color, shape, and texture features are common features for images, while kinetic energy, motion vectors are used to describe video data. Correspondingly, a query can be also represented in terms of features so that it can be matched against the data.

**Query-by-Example (QBE):** A method of forming queries that contain one or more media object(s) as examples with the intention of finding similar objects. A typical example of QBE is the function of "See Similar Pages" provided in the Google search engine, which supports finding web pages

similar to a given page. Using an image to search for visually similar images is another good example.

**Multi-modality:** Multiple types of media data, or multiple aspects of a data item. Its emphasis is on the existence of more than one type (aspects) of data. For example, a clip of digital broadcast news video has multiple modalities, include the audio, video frames, closed-caption (text), etc.

**Cross-Media Retrieval:** As an extension of traditional multimedia retrieval methods, Cross-Media retrieval can be regarded as a unified multimedia retrieval approach which tries to breakthrough the modality of different media objects. For example, when user submits a "tiger" images, the system will return some "tiger"-related media objects with different modalities, such as the sound tiger roars and the video describing a tiger is capturing animals, etc.