2009 IEEE International Conference on Data Mining Workshops

MSRA-MM 2.0: A Large-Scale Web Multimedia Dataset

Hao Li*
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, 100190, China
lihao@ict.ac.cn

Meng Wang
Microsoft Research Asia
Beijing, 100080, China
mengwang@microsoft.com

Xian-Sheng Hua
Microsoft Research Asia
Beijing, 100080, China
xshua@microsoft.com

***Abstract*—In this paper, we introduce the second version of Microsoft Research Asia Multimedia (MSRA-MM), a dataset that aims to facilitate research in multimedia information retrieval and related areas. The images and videos in the dataset are collected from a commercial search eingine with more than 1000 queries. It contains about 1 million images and 20,000 videos. We also provide the surrounding texts that are obtained from more than 1 million web pages. The images and videos have been comprehensively annotated, including their relevance levels to corresponding queries, semantic concepts of images, and category and quality information of videos. We define six standard tasks on the dataset: (1) image search reranking; (2) image annotation; (3) query-by-example image search; (4) video search reranking; (5) video categorization; and (6) video quality assessment.**

***Keywords*-benchmarking dataset; reranking; annotation;**

I. Introduction

The last decade has witnessed great advances in multimedia information retrieval. An encouraging phenomenon is that research and industrial communities are getting closer in this area [1]. First, search engines are becoming popular tools for collecting multimedia data: large-scale web data can be easily obtained by collecting the search results of various queries. Some efforts have been conducted on efficient web data mining and management [19]. Second, many research works have been conducted on refining the search result, such as reranking [22][28][24], and these approaches can be easily implemented on the current search engines.

However, gap still exists between the two communities. For example, most research works on image and video search adopt simple object- or scene-level queries such as "sunset", "bird" and "car", whereas real users' queries for image and video search engines are often much more complex, such as a movie and a rock star. In addition, despite large-scale multimedia search has been promoted for a long time in research community, most research experiments are still conducted on limited data, such as tens of queries for ranking and reranking, which can hardly verify the robustness and practical usefulness of the algorithms.

A carefully designed large-scale dataset is highly desired in order to bridge this gap. The dataset needs not only to provide benchmarking data for developing and evaluating algorithms for multimedia search as well as connect them with state-of-the-art algorithms and results in industrial community. Currently, there are many different publicly available datasets served as useful resources in computer vision and multimedia communities. However, it lacks a dataset that is particularly designed for web image and video search. Microsoft Research Asia Multimedia (MSRA-MM) is a dataset that is intended to facilitate research in image and video search via open and metrics-based evaluation. The first version of MSRA-MM [27] was released at early March, 2009 as a prototype, and here we introduce an advanced version – MSRA-MM 2.0. In this version, we enlarged the data scale and re-organized the structure of the dataset. Analogous to the previous version, MSRA-MM 2.0 also contains two sub-datasets, i.e., an image dataset and a video dataset, that are collected from a commercial search eingine. The image part contains 1,011,738 images that are collected from 1165 queries and the video part contains 23,517 videos that are collected from 217 queries. The associated web pages are also downloaded and surrounding texts are extracted. We provide low-level features extracted from images and video key-frames as well as the annotation results on their relevance, semantic concepts, categories and qualities. Based on the above information, different tasks can be investigated on the dataset, including image/video search reranking, annotation, query-by-example image search, video categorization and quality assessment.

As mentioned in [11], an ideal benchmark dataset should have following requirements: (1) the dataset set should be *representative* of an interesting image retrieval area; (2) The *ground truths* should be available for the dataset so that objective evaluations can be performed; (3) The dataset should be easily *accessible* and freely *redistributable*; (4) it is important to have a set of *standardized* tests associated with the database. We have designed the MSRA-MM dataset following these guidelines. First, the images and videos contained in MSRA-MM are collected with top queries of

*This work was performed when Hao Li was visiting Microsoft Research Asia as a research intern.

[1]The appearance of MultiMedia Grand Challenge (MMGC), which is associated with ACM Multimedia 2009, further confirms this trend. It presents a set of tasks that are designated by Google, Yahoo, HP, etc. to researchers. It indicates the problems that industrial leader are interested in multimedia.

978-0-7695-3902-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ICDMW.2009.46

Table I
A COMPARISON OF DIFFERENT IMAGE DATASETS ("-" MEANS THE TERMS ARE NOT DESCRIBED IN THE RELATED PAPERS OR TECHNICAL REPORTS).

	Caltech-256	**LabelMe**	**PASCAL'08**	**Lotus Hill**	**TinyImage**	**ImageNet**	**NUS-WIDE**	**MIR Flickr**	**MSRA-MM**
Released Year	2006	2005	2008	2007	2008	2009	2008	2008	2009
Institute	Caltech	MIT	Oxford	LHI	MIT	Princeton	NUS	Leiden	MSRA
Image Source	Google, Pic-Search	Personal	Flickr	Personal	Search Engines	Search Engines	Flickr	Flickr	Search Engine
Image Number	30,607	163,054	10,057	636,748	79,302,017	3,200,000	269,648	25,000	1,011,738
Image Quality	Varied	-	-	-	Low	High	Varied	High	Varied
Public Availability	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No
Duplication Check	Yes	-	-	-	-	-	Yes	-	No
Category Number	256	No	20	268	75,062	5,247	No	No	1165
Image/Category	80∼827	No	48∼1025	-	≈1000	500∼1000	No	1∼845	511∼943
Features	No	No	No	No	No	No	Yes	No	Yes
Annotation Type	-	Polygon	Bounding box	Bounding box; Sketch	-	Concept ground truth	Concept ground truth	Concept ground truth	Relevance; Concept ground truth;

a commercial search eingine, which can reflect real users' needs. Second, the relevance of each image and video with respect to the corresponding query is labeled. We have also provided the ground truths of semantic concepts, categories and qualities. Third, although we cannot distribute original images, videos and web pages due to copyright issues, we have shared the features of image and video key-frames as well as the ground truths. Finally, we have defined a set of tasks on the dataset. We will provide the split of training set and testing set for these tasks as well as the baseline results for each task later.

The organization of the rest of this paper is as follows. In Section II, we provide a short review on the related datasets. In Section III, we describe the construction process of MSRA-MM. Section IV introduces the five tasks that can be investigated on the dataset.

II. RELATED DATASETS

A. Image Datasets

There are many well-known image datasets in computer vision and multimedia communities. Corel [15], a widely used dataset that contains more than 800 photo CDs, has greatly facilitated research in image classification, annotation and search. However, despite the dataset contains a huge number of photos, evaluations are often only done on small subsets (such as Corel2000 [17] for image annotation and Corel5000 [7] for image categorization and query-by-example). Several datasets are designed to facilitate research in object recognition and detection, including Caltech-101/256 [6], LabelMe [18] and Pascal VOC. With the advances in storage and computation devices, larger datasets are emerging to cover more image classes and pose challenges to algorithms in handling large-scale multimedia analysis and search. TinyImage [23] consists of about 80 million low resolution images which are collected by performing image search with regarding each noun term in WordNet [5] as a query. It addresses the problem that how large a dataset needs to be when simple k-NN algorithm is sufficient to perform robustly. ImageNet [4] provides a well-structured, large-scale, accurate and diverse image database that is closely integrated with WordNet. Though most of these dataset are collected from the Web, they have not kept the meta-data of the original data, such as the names, tags and surrounding texts of images, that can be valuable resource for web multimedia search. Some other datasets get labeled data freely from Flickr, a well-known social media website. NUS-WIDE [3] comprises over 269,000 images and 5,000 user-provided tags from Flick and the ground truths of 81 semantic concepts are provided on these images. It also provides different low-level features for each image. MIR Flickr [11] consists of 25,000 high-quality images that are also collected from Flickr with attribution lesions that allow research redistribution. However, these two datasets have not kept the ranking information of images, and thus we cannot conduct research on ranking and reranking on the data. In addition, the associated tags are noisy and the concepts of images are not exhaustively labeled. Table I illustrates the information of these datasets as well as their comparison in the following aspects:

- Image Source. It describes where the images are collected from, such as search engines, social media websites or personal photos.
- Image Quality. It gives a general description of image qualities, e.g., only high quality images are kept, or images with varied qualities are contained.
- Public Availability. It indicates whether the source images are publicly available.
- Duplication Check. It indicates that whether the dataset removes duplicate images.
- Features. It means whether features are extracted and

Table II
A COMPARISON OF DIFFERENT VIDEO DATASETS.

	TRECVID'06	TRECVID'07	TRECVID'08	Kodak Consumer Video	MSRA-MM Video
Released Year	2006	2007	2008	2008	2009
Institute	NSIT	NIST	NIST	Kodak Research	MSRA
Hours	337	109	218	-	1,336
Videos	536	219	438	3,231	23,517
Shots	169,156	36,262	72,028	No	1,041,034
Key-frames	298,158	43,616	86,000	5,166	1,041,034
Video Categories	News video	News video, documentary, educational programming, archival video		Home video and web video	Web video
Annotation	Concepts, Shot boundaries, Relevance(concepts and relevance are labeled in shot-level)			Concepts(labeled in video-level for web videos and labeled on key-frames for home videos)	Shot boundary, Relevance, category, quality(relevance, category and quality are labeled in video-level)
Metadata	Automatic speech recognition and machine translation transcripts			Video URL, Tags, Category	Surrounding Text
Concepts	20	20	20	25	No
Search Queries	24	24	48	No	217

shared for the images.

- Annotation Type. It means what kinds of annotation are provided, such as the labeling of object bounding boxes and the concept of images.

B. Video Datasets

In comparison with image datasets, video datasets are fewer due to the large volume of data. TRECVID [20] organized by NIST can be regarded as the de facto benchmark in video search and the related areas. It provides a large video collection as well as uniform evaluation procedures for researchers to compare their results. It supports multiple tasks, including shot boundary detection, high-level feature extraction, search, copy detection, etc. However, a problem is that the videos used in TRECVID are mainly news videos, and only from 2007 it begins to incorporate more video genres such as documentaries, educational programming and archival videos. Kodak consumer video benchmark [12] consists of 1358 videos that are collected from 100 users and 1873 videos that are downloaded from YouTube. A lexicon of 25 concepts is constructed and all videos are annotated with these concepts.

Different with the TRECVID and Kodak consumer video datasets, the video data in MSRA-MM are collected from a commercial search eingine with top queries and we have kept their metadata and ranking information. The videos cover a wide range of genres and topics. Therefore, it can be used in more applications, such as the research on video ranking and reranking. Table II illustrates the information of TRECVID 2006, TRECVID 2007, TRECVID 2008, Kodak consumer video dataset and MSRA-MM as well as their comparison in the following aspects:

- Video Categories. It describes the genres of video, such as broadcast news, home video or web video. It is worth noting that here web videos indicate the videos that are collected from the web and they are actually heterogeneous, as they may contain sports videos, movies, game videos, etc.
- Annotation. It describes in what aspects the videos are labeled, such as shot boundary, relevance to query (topic), and concept ground truths.
- Metadata. It lists the type of metadata associated with videos, such as user tags, web page surrounding text.
- Concepts. It indicates the number of concepts that are defined if video concept detection is supported on the dataset.
- Search queries. It indicates the number of queries that are defined on the dataset if video search is supported on the dataset.

III. THE CONSTRUCTION OF MSRA-MM

A. Image Dataset

1) Query Selection: We obtain the query log of a commercial search eingine on Jan 6th 2009 which contains 1,048,576 queries and their frequency information. Then we select 1009 queries that have frequencies above 1000. These queries are manually categorized into 8 categories, i.e., Animal, Cartoon, Event, Object, Scene, PeopleRelated, NamedPerson, and Misc[2]. In addition, we further add 88 person names from the 2008 TIME 100 in order to cover more persons[3]. The information of the categories is listed in Table III. For each query, the retrieved images are collected together with their thumbnails and corresponding web pages.

[2]Misc mainly refers to abstract concepts (e.g., love), trademarks (e.g., puma), movie or TV shows (e.g., twilight, Hanna Montana) and place names (e.g., Africa). These queries are usually highly ambiguous and can be understood in different aspects.

[3]We observe that named persons occupy a large proportion in the original query log. However, many of them are adult celebrities and we need to filter them out. Thus we seek other ways to add other named persons in order to balance the distribution of query types.

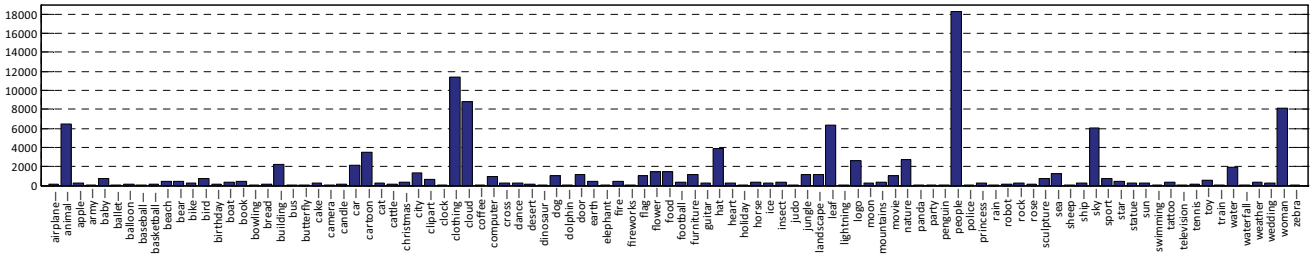

Figure 1. The Numbers of Positive Images for 100 Concepts.

Table III
THE STATISTICAL INFORMATION ABOUT IMAGE DATASET.

Category	Query	Image	Image/Query		Exemplary queries
			Max	Min	
V1.0	68	60,257	923	827	angles, baby, cake, fish
Animal	100	88,241	934	656	alligator, bat, cattle
Cartoon	92	77,447	910	700	air gear, final fantacy
Misc	288	249,440	950	673	japan, titanic, adidas
Event	78	69,493	930	681	olympic, wedding, wwe
Object	295	257,060	943	719	airplane,bed,toy
People	68	59,082	929	624	girl, snowman, baby
Person	40	33,245	893	741	tom hanks, will smith
Scene	48	43,121	939	858	desert, rainbow
TIME08	88	74,352	930	511	barack obama, steve jobs
Total	1,165	1,011,738	943	511	

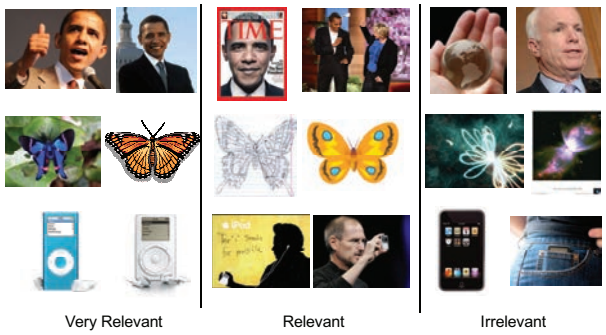

Figure 2. Several Exemplary Images with Different Relevance Levels with Respect To "barak obama", "butterfly" and "ipod".

The ranking information and other metadata such as image URLs and web page URLs are all recorded in XML files.

2) Feature Extraction: The 7 global features we used include: (1) 225D block-wise color moment [21]; (2) 64D HSV color histogram [9]; (3) 144D color correlogram [10]; (4) 256D RGB color histogram [6]; (5) 75D edge distribution histogram [16]; (6) 128D wavelet texture [13]; and (7) 7D face features. All images and key-frames are resized to a fixed width of 240 pixels before extraction. The information about these features is listed in Table IV.

3) Surrounding Text Extraction: We provide surrounding text in the form of term frequency. For each web page, the VIPS [2] algorithm is adopted to segment the web pages into blocks, then the texts in the block that contains the image's or video's URL are extracted as surrounding texts. The texts are split into single words and only those that are nouns in the WordNet are kept with their frequencies. We extract 66805 unique nouns and assign numerical IDs to them in alphabet sequence. Then the surrounding texts can be represented as a list of IDs and frequencies.

4) Image Annotation:

a) Semantic Concepts: We construct a lexicon of 100 concepts that are selected from the 1165 queries. We randomly select 50,000 images from the 1 million images to label the concept ground truths. For each image, it is manually labeled as "positive" or "negative" with respect to every concept. Figure 1 illustrates the number of positive samples for each concept.

b) Relevance: For each image, its relevance with respect to the corresponding query is manually labeled with three levels: very relevant, relevant and irrelevant. These three levels are indicated by scores 2, 1 and 0, respectively. Each query has been assigned a description before the manual labeling. Several ambiguous queries may have more than one meaning. For example, "apple" may refer to fruit, computer and mobile phone. In our work, the images corresponding to different meanings are all regarded as relevant or very relevant. Figure 2 illustrates several exemplary images of "barak obama", "butterfly" and "ipod" with different relevance levels.

B. Video Dataset

In comparison with the version 1.0, we have added 52 top queries from the query log and collected 13,240 videos accordingly. For each video, we perform shot boundary detection using the algorithm in [8] and a key-frame is selected from each shot. The features illustrated in Table IV are extracted from each key-frame. The surrounding text information is extracted using the method introduced in Section 3.1.3. The information about the videos can be found in Table II.

Analogous to images, the relevance of each video with respect to the corresponding query is manually labeled. In addition, we also label the category and quality information of videos. For each video, it is labeled as whether belong to

Table IV
THE DESCRIPTION OF LOW-LEVEL FEATURES

Feature Name	Dim	Description
Block-wise Color Moment	225	Each image is split into 5-by-5 blocks, and 9-dimensional color moment features are extracted from each block.
HSV Color Histogram	64	64-dimensional histogram features extracted in HSV color space.
Color Autocorrelogram	144	HSV color components are quantized into 36 bins with 4 different pixel pair distance k, i.e., k = 1, 3, 5, 7.
RGB Color Histogram	256	256-dimensional histogram features extracted in RGB color space
Edge Distribution Histogram	75	The image is divided into 5 blocks and 15-dimensional EDH features are extracted.
Wavelet Texture	128	Wavelet transform is performed on each image with recursive filtering and sub-sampling, and 128-dimensional features are extracted using the mean and standard deviation of the energy distribution of each sub-band at different levels.
Face	7	The features include the number of faces, the ratio of face area and the position of the largest face.

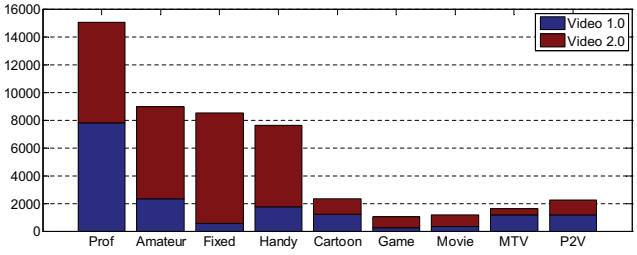

Figure 3. The Distribution Information of Video Categories.

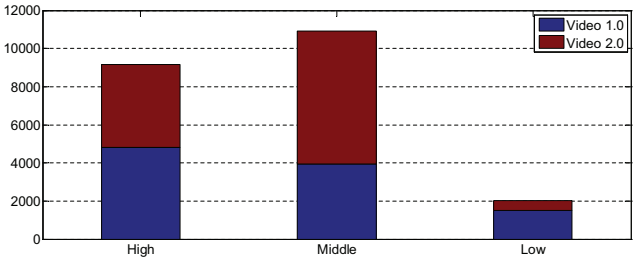

Figure 4. The Distribution Information of Video Qualities.

the following nine non-exclusive categories: (1) professional videos; (2) amateur videos; (3) handy videos; (4) fixed videos; (5) movie videos; (6) cartoon videos; (7) game videos; (8) MTV videos; and (9) Photo2Video (P2V) videos. Therefore, each video has nine category flag values. The distribution of category can be seen in Figure 3.

The quality of each video is manually labeled with three levels: high quality (score 1), middle quality (score 0) and low quality (score -1). The distribution of different video qualities is illustrated in Figure 4.

IV. TASKS

A. Image/ Video Search Reranking

Content-based image/video search reranking is a technique that aims to adjust the ranking lists obtained based on textual information by exploring visual content, such that better ranking lists can be obtained. We have kept the original order of the crawled images and videos in MSRA-MM. Therefore, the performance of the image search engine can be directly measured, and this can be considered as the baseline which indicates the performance of an industrial search engine. We can conduct research on reranking on the dataset and the performance can directly be compared against industrial results.

B. Image Annotation

Based on the 50,000 images that are labeled with 100 concepts, we can conduct research on image annotation. Different from many image benchmark datasets that only support categorization, such as Caltech 101 and Caltech 256, our concepts are non-exclusive, i.e., the concepts may co-occur in an image, and thus we can investigate multi-label annotation techniques on the dataset [26][25]. Because the numbers of positive samples are usually much less than negative samples in the task, the classification accuracy is not a preferred performance measure. Therefore, we will adopt Average Precision (AP) as the performance evaluation metric, which actually measures the performance of ranking that is generated with the relevance scores of images.

C. Query-By-Example Image Search

We select an example image for each of the 100 concept queries. With the features and ground-truth of query images and the 50,000 images, we can conduct research on query-by-example image search, including relevance feedback. Figure 5 illustrates the 100 example images.

D. Video Classification

We can investigate video classification [1] techniques based on the provided video category information. For example, we can classify the videos into professional videos and amateur videos and classify amateur videos into fixed videos and handy videos. We can also perform detections for certain categories, such as P2V video detection.

E. Video Quality Assessment

Web video quality assessment [14] is a topic that receives less attention, but it is actually useful in video search. For example, we can filter out low-quality videos in the search results or perform reranking to prioritize high-quality videos in ranking lists. Based on the provided quality information, we can conduct study on automatic video quality assessment.

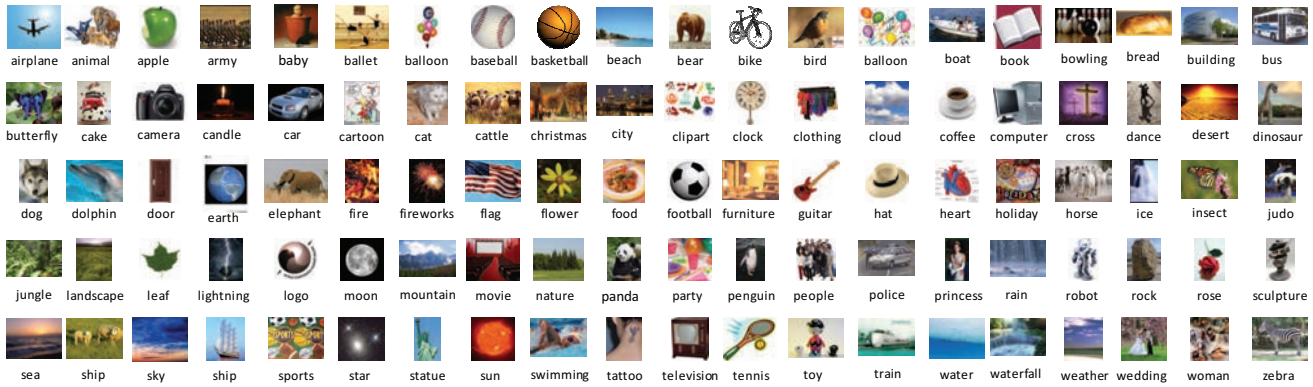

Figure 5. 100 Query Images for QBE Task.

V. Conclusion and Future Work

In this paper, we introduced MSRA-MM 2.0, a new large-scale web multimedia dataset that contains about 1 million images and 20,000 videos. This dataset allows researchers to directly compare their algorithms with industrial results. We identify six tasks on the dataset: (1) image search reranking; (2) image annotation; (3) query-by-example image search; (4) video search reranking; (5) video categorization; and (6) video quality assessment. In the future work, we will provide baseline results for the six tasks. We will seek more efficient ways for data sharing and collaborative annotation.

References

[1] D. Brezeale and D. J. Cook. Automatic video classification: a survey of the literature. *Trans. on SMC*, 2008.

[2] D. Cai, S. Yu, J. R. Wen, and W.-Y. Ma. Block-based web search. *SIGIR*, 2004.

[3] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. NUS-WIDE: A real-world web image database from national university of singapore. *Proc. of International Conference on Image and Video Retrieval*, 2009.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. *CVPR Workshops*, pages 248–255, 2009.

[5] C. Fellbaum. *WordNet: An Electronic Lexical Database*. 1998.

[6] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, Caltech, 2007.

[7] N. Herve and N. Boujemaa. Image annotation : which approach for realistic databases? *Proc. of International Conference on Image and Video Retrieval*, 2007.

[8] X.-S. Hua, L. Lu, and H.-J. Zhang. Robust learning-based TV commercial detection. *ICME*, 2005.

[9] J. Huang, S. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlogram. *CVPR*, 1997.

[10] J. Huang, S. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlogram. *CVPR*, 1997.

[11] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. *Proc. of International Conference on Multimedia Information Retrieval*, 2006.

[12] A. Loui, J. Luo, S.-F. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, and A. Yanagawa. Kodak's consumer video benchmark data set: concept definition and annotation. *Proc. of International Conference on Multimedia Information Retrieval*, 2007.

[13] B. S. Manjunath and W.-Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. on PAMI*, 1996.

[14] T. Mei, X.-S. Hua, C. Z. Zhu, H. Q. Zhou, and S. Li. Home video visual quality assessment with spatiotemporal factors. *Trans. on CSVT*, 2007.

[15] H. Müller, S. Marchand-Maillet, and T. Pun. The truth about corel - evaluation in image retrieval. *Proc. of International Conference on Image and Video Retrieval*, 2002.

[16] D. K. Park, Y. S. Jeon, and C. S. Won. Efficient use of local edge histogram descriptor. *Proc. of ACM International Conference on Multimedia*, 2000.

[17] Y. Rui and G. J. Qi. Learning concepts by modeling relationships. *MCAM*, 2007.

[18] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 2008.

[19] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. *Proc. of International Conference on Computer Vision*, 2007.

[20] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. *Proc. of International Conference on Multimedia Information Retrieval*, 2006.

[21] M. Stricker and M. Orengo. Similarity of color images. pages 381–392, 1995.

[22] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. *Proc. of ACM international Conference on Multimedia*, 2008.

[23] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *IEEE Trans. on PAMI*, 2008.

[24] L. S. K. W. H. Hsu and S. F. Chang. Video search reranking via information bottleneck principle. *Proc. of ACM International Conference on Multimedia*, 2006.

[25] M. Wang, X. S. Hua, R. C. Hong, J. H. Tang, G. J. Qi, and Y. Song. Unified video annotation via multi-graph learning. *IEEE Trans. on Circuits and Systems for Video Technology*, 19(5), 2009.

[26] M. Wang, X.-S. Hua, J. H. Tang, and R. C. Hong. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *IEEE Trans. on Multimedia*, 11(3), 2009.

[27] M. Wang, L. J. Yang, and X. S. Hua. MSRA-MM: Bridging research and industrial societies for multimedia information retrieval. Technical report, Microsoft, 2009.

[28] R. Yan, A. G. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. *Proc. of International Conference on Image and Video Retrieval*, 2003.