WIKI-CMR: A WEB CROSS MODALITY DATASET FOR STUDYING AND EVALUATION OF CROSS MODALITY RETRIEVAL MODELS

Wei Xiong[1], *Shuhui Wang*[2], *Chunjie Zhang*[1], *Qingming Huang*[1,2]

Schoold of Computer and Control Engineering, University Of Chinese Academy Of Science[1]
Key Lab of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences[2]
{wxiong, shwang, cjzhang, qmhuang}@jdl.ac.cn

ABSTRACT

With the popularity of Web multimedia data, cross-modality retrieval becomes an urgent and challenging problem. Bridging the semantic gap between different modalities and dealing with abundant data are the main challenges for cross-modality retrieval. A well-designed dataset could provide a platform for developing the state-of-the-art cross-modality retrieval algorithms. However, existing Web cross-modality datasets are small in size, or do not contain the full information, for example, the hyperlink structure. In this paper, we introduce a new Web cross-modality dataset called "WIKI-CMR" by selecting Wikipedia as the reliable and information-rich data resource, and collect data with a smart crawling strategy. This dataset is comprised of 74961 documents with textual paragraphs, images and hyperlinks. All documents are categorized into 11 semantic topics. We point out several challenges on this dataset and use this dataset to evaluate some well-known cross-modality retrieval models.

Index Terms— Multimedia, dataset, retrieval, cross-modality

1. INTRODUCTION

Multimedia information retrieval is a grand challenge in computer science. The aim is to make capturing, storing, finding and using digital media an everyday occurrence in our computer environment [1]. With the spread of social media such as Twitter, Facebook, Flickr, and YouTube, information is no longer delivered in a single modality as before. The amount of multi-modality data grows explosively every day. As multi-modality data can represent a certain topic in a more vivid way, they are produced, propagated and shared among common Web users. At the same time, the demand for searching other modality data by providing a single modality query becomes stronger and stronger. For example, when we search information of "batman", we wish to get a brief literal introduction, some correlated images and some short videos. While mining semantic relation among different data modalities is quite challenging due to the heterogeneity and complexity of multi-modality data. Therefore, cross-modality retrieval becomes a challenging problem to be solved.

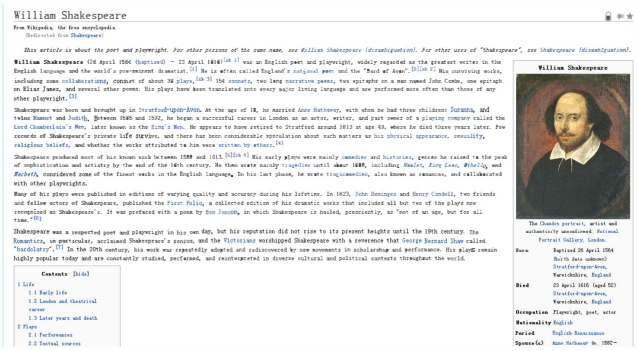

Fig. 1. an example of Wikipedia webpage

During the last decades, different algorithms are proposed for cross-modality retrieval. In the context text-to-image retrieval, canonical correlation analysis (CCA) is a classical but still efficient method which learns the subspaces of two data modalities that maximize their correlation. In [2], Hardoon *et al.* propose kernel canonical correlation analysis (KCCA) to deal with cross-modality retrieval tasks especially for mate-retrieval. By projecting cross-modal data into an implicit and non-linear space using the kernel trick, KCCA improves the performance of cross-modality retrieval. In [3], sparse canonical correlation analysis (SCCA) is proposed to deal with the high data dimensionality, and learns a set of sparse projection vectors. SCCA achieves better performance than KCCA when the number of original features is larger than the number of cross-modal data pairs, by learning the partial correlation model between the two data modalities. Rasiwasia *et al.* [4] combine CCA with semantic abstraction for cross modality retrieval, and the accuracy of cross-modality retrieval has been improved in terms of topic classification accuracy. To use the unlabeled data to mitigate over-fitting and enhance

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61025011, in part by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), in part by China Postdoctoral Science Foundation: 2012M520434. .

the model capacity, Semi-CCA [5] is introduced to image annotation task, and is proved to be more robust than CCA.

Besides CCA, the latent topic model such as latent Dirichlet Allocation (LDA) is another solution to cross-modality retrieval. In [6], Blei *et al.* introduce a new approach, Corr-LDA, for modeling annotated data with multiple types which can be applied to cross-modality retrieval. Based on Corr-LDA research, Jia *et al.* [7] construct a Markov random field over LDA topic models which does not requires strict one-to-one correspondence between data modalities, thus better effectiveness and flexibility are achieved for cross-modality retrieval. By embedding two spaces to a latent space, the above-mentioned algorithms can measure the relation of different data modalities in determined or probabilistic manner. Besides, Zhuang *et al.* [8] proposed a new similarity measurement of different data modalities by constructing an uniform cross-media correlation graph. Tang *et al.* [9] focus on the relevance of different modalities in Web image annotation. And Geng *et al.* [10] introduce a method of combining information of different modalities in Web image research task.

Although remarkable success has been achieved, we notice from above-mentioned studies that they established their studies on different small scale datasets. For example, in [2], they use a combined Web image-text dataset established by [11] with 1200 data items that are categorized into 3 categories. In [3], they use europal German-Danish paired bilingual corpus with only 150 documents to evaluate SCCA. A dataset with 1000 images, 300 audio files and 720 texts collected from different webpages is used in [8]. Compared with the Internet data corpus, the data sizes in these existing datasets [2] [3] [8] [11] are too small, hence the reported results are not strong and statistical significant.

When facing with more complicated cross-modality retrieval problems, previous algorithms may fail and underfitting may be incurred because the model can not make use of more information, such as the topic category information and hyperlink structure. Rasiwasia *et al.* [4] start with the pioneering research of Web cross-modality retrieval by constructing correlation model on the dataset of Wikipedia's "featured article". consequent research, such as [7] also collect data from Wikipedia in a similar manner. Although the category information has been used in their study, the hyperlink among webpages is missed in a naive pre-processing procedure of collecting the dataset, which does not well reflect the true data structure from the Web corpus. Besides, a well-known dataset "wikidata" introduced by imageclef competition does not contain enough literal information.

For the research of Web cross-modality retrieval, Wikipedia is a reliable data source for constructing a web cross-modality dataset as it provides well-edited multi-modal data, and it topic coverage is very diversified. Besides, unlike some websites with amount of useless and disordered links, Wikipedia provides hyperlinks that normally link two semantic highly related webpages. With hyperlinks, webpages in Wikipedia are highly structured. In this paper, we design a smart data crawling strategy to collect the data and construct the dataset to ensure the wide topic coverage and preserve the useful hyperlink of the content.

The rest of paper is organized as follows: we explain how we construct the dataset in section 2. We illustrate the properties of our dataset in section 3. In section 4, we point out some challenges on this dataset. In section 5, we conduct comparison on some cross-modality retrieval algorithms. Future work is provided in section 6.

2. DATASET CONSTRUCTION

2.1. Collecting and processing data

Generally, a webpage of Wikipedia usually includes three types of information, the literal information presented as a text, visual information presented as images and content interrelationship presented as hyperlinks. Therefore, when collecting data we stored the webpage information in three subfolders as text, images and hyperlinks respectively.(Our collecting strategy compliance with http://en.wikipedia.org/robots.txt)

Instead of randomly choosing the content, we consciously design a list of 500 items which are mainly focusing on the fields of culture, geography, natural, people and history. Such items offer more images and reliable semantic information compared with the items from other fields such as mathematic, technology and philosophy. With the downloaded 500 webpages, we have a considerable number of hyperlinks from them which can be used as new items. After collecting those new items' webpage information, we will have more hyperlinks again that can be used as new items. This breadth-first strategy can guarantee that we can find at least one linked webpage for every webpage as explained in section 2.3.

After 3 month collection, we obtain 6381 webpages. All the webpages are labeled into 11 categories (*culture and the arts, people and self, geography and places, society and social science, history and events, natural and physical science, technology and applied science, religion and belief system, health and fitness, mathematic and logic, philosophy and thinking*) [12] by several non-expert students. The score ranging from 0 to 2 indicates the confidence of correlation between the webpage and the category. Score 2 denotes that the webpage is categorized into this category. One webpage can be and only be categorized into one category. Score 1 represents the webpage is semantically relevant with the category but the webpage doesn't belong to this category. Otherwise, the score 0 means the webpage has no semantic relationship with the category. At last, we have a label matrix for 6381 webpages.

The texts are split into several paragraphs according to the headline information located at the middle top of webpages, as showed in Fig.1. In fact, the text of *William Shakespeare* webpage is split into 26 paragraphs based on the content blocks in our data processing procedure. Some of the paragraphs are eliminated during the feature extraction pro-

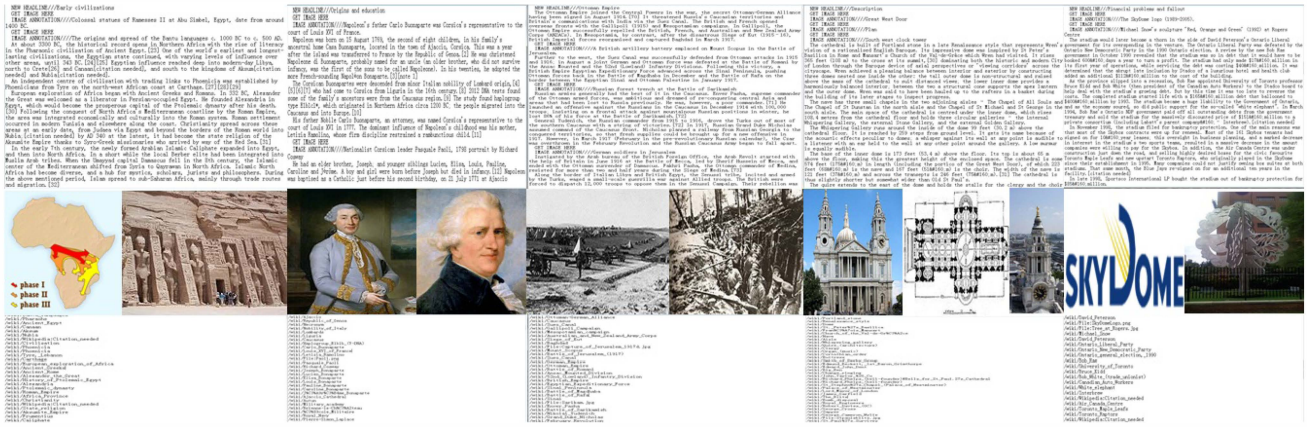

Fig. 2. some examples of our dataset

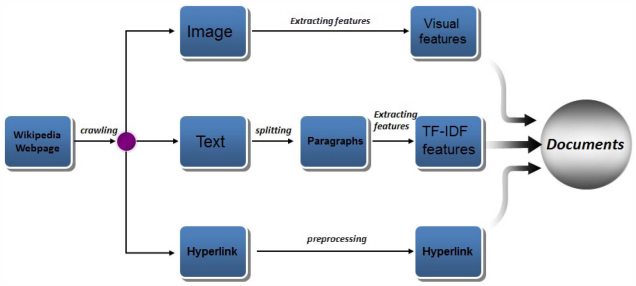

Fig. 3. collecting and preprocessing data

cess discussed in section 2.2. We also record how the image is original placed with some paragraph, which is the important co-occurrence information of two data modality. Each paragraph may have several related images or none. And those hyperlinks within the webpage that do not link to any Wikipedia content webpages are eliminated. After the procedure, the dataset contains 74961 documents. Each document includes one paragraph, one related images(or no image), category label, and hyperlinks.

2.2. Feature extraction

Since the Web visual content is usually divergent in semantics (our dataset cover all the 11 diversified topic categories), different visual features are needed to provide complementary visual description. To this end, we can extract some state-of-the-art visual features. Dense sift feature [13] shows significantly improved performance on the challenging scene categorizing task. We extract dense SIFT descriptors and to learn a codebook with size of 512 by sparse coding [14]. Each image is represented by a 10754-dimensional vector with the max-pooling and 3-level spatial pyramid method [15]. Gist feature [16] is used for scene categorizing and provides an efficient way for visual context modeling. We calculate two types of GIST feature in our dataset namely, the GIST on the whole image and the concatenation of GIST descriptors of 4*4 spatial image blocks. LBP feature [17] is a theoretically and computationally simple yet efficient feature for texture modeling and face recognition. Due to the large number of texture and human face images in our dataset, LBP feature is also calculated for each image. Meanwhile, PHOG feature [18] significantly outperforms existing features on object detection. 180 degree and 360 degree of PHOG are calculated for each image respectively. Self-similarity [19] is another well-known feature for measuring similarity between visual entities (images or videos). The coding and pooling method of Self-similarity is the same as Dense-Sift.

We use the Term Frequency-Inverted Document Frequency (TF-IDF), a very common but efficient feature, to represent each textual paragraph. We first construct a codebook of more than 380k words from all the paragraphs. Then the refined codebook of 73212 words is generated by removing stop words and words with extremely low and high frequencies. However, we find that the text feature with 73212 dimensions will lead to explosive computation burden and poor performance brought by *the curse of dimensionality*. Therefore, we use SVD decomposition to reduce the feature dimension to 2000. Finally, each of the 74961 paragraphs is represented as a 2000-dimensional feature vector.

2.3. dataset organization

Each paragraph has a category label which is the same as the original webpage's category label. As the hyperlinks link a webpage to another webpage, it is difficult to assign the hyperlinks to paragraph-to-paragraph level. Hence, all the documents have the same hyperlink list as the original webpage . In this dataset, 23490 documents contain a paragraph presented as a 2000-dimensional TF-IDF text feature, correlated images represented as 8 types of visual features, label information and hyperlinks, and the rest of documents only con-

tain a paragraph, label information and hyperlinks due to the lack of images in the original webpage. And a comparison of existing cross-modality datasets are conducted in Table. 1 [1].

3. PROPERTIES OF THE DATASET

Reliability Wikipedia is a free, collaboratively edited and multilingual Internet encyclopedia. The content in Wikipedia is reliable and rich. Articles and images in the same webpage are highly semantically related. It assures that most co-occurred texts and images can be used as ground-truth without complex processing. Moreover, the Wikipedia webpages are edited by millions of volunteers, anyone can refine the webpage if they find mistakes or want to enrich the webpage. Therefore, the updating and modifying speed of Wikipedia webpage is faster than normal webpages. Some investigation [20] showed that the quality of the articles from Wikipedia came close to the level of accuracy of Encyclopedia Britannica and had a similar rate of "serious errors".

Imbalance. Different from other datasets used in previous research, this dataset is imbalanced. Images and paragraphs are not strictly organized into one-to-one correspondence. Some paragraphs may have more than one image while some don't have a co-occurred image. Single modality data will occur in our dataset while they are usually eliminated in other datasets. This imbalanced dataset better reflects the true data infrastructure of the Internet. Moreover, people would be likely to conduct cross-modality retrieval among different websites. For example, they may search textual descriptions from Twitter by providing an image from Flickr. In such situation, the imbalance property is ubiquitous and one should develop better cross-modality retrieval algorithms to deal with this difficulty.

Hyperlink. When an Web ontology appears in a paragraph, a hyperlink is produced in the form of blue words (as shown in Fig.1). Hyperlinks reflect the semantic relation between ontologies of the webpages (documents) and also represent the interaction, collaboration, or influence information among webpages (documents). The information of intra-relationship among documents has never been introduced to cross-modality retrieval dataset while it contains much information both in semantic and webpage link structure. These intra-relationship information could help the algorithms to learn the semantic relation among documents better.

4. CHALLENGES OF THE DATASET

4.1. Effective retrieval models

For Web cross-modality retrieval, Image-to-text and text-to-image retrieval are fundamental yet challenging problems that need further investigation. The main challenges on this dataset include: 1) how to model the intra-modal relation and inter-modal relation among the cross modal documents; 2) how to take advantage of various types of information for model development, including the intra-modal relation, inter-modal relation, label information and structure information; 3) how to provide a unified measurable representation on which the semantic relevance of cross modal data can be calculated. Besides, another interesting challenge on this cross modal dataset refers to the intra-model retrieval. For example, given an image query, one may like to retrieve those semantically relevant images. In this case, the rich context such as the hyperlinks and co-occurred textual description can be very useful for model construction.

4.2. Dealing with hyperlinks

Link analysis was firstly used for ranking webpages. Generally speaking, a webpage with more hyperlinks is assumed to be more important. Such phenomenon illustrates that hyperlinks contain much information that help us improve the cross-modality retrieval performance. Besides, link analysis draws much attention as it can be used to mining implicit semantic relevance among Web data. In documents taxonomies context, Ho *et al.* [21] shows that automatically induced taxonomies can be made more robust by combining text with relational links. In our dataset, hyperlink is only produced when one taxonomy appears in another's article. Therefore, analyzing hyperlink help us better recover and reconstruct the semantic relevance structure of the dataset.

Moreover, hyperlink can be used as a special similarity measurement among documents. The importance of the hyperlinks may be diversified due to their locations and frequency. Once the importance of the hyperlinks are properly weighted, the semantic similarity model among documents can be constructed, which provides a potential possibility that the accuracy of cross-modality retrieval can be improved.

5. EXPERIMENTS AND ANALYSIS

In this section we use this dataset to evaluate some cross-modality retrieving models. As CCA and its variants can not deal with unpaired data, we choose 23490 multimedia documents that contain a paragraph and an image to evaluate these models. Each document will contain only one image after a random sampling procedure.

5.1. CCA

In this section, we conduct our experiments with original CCA. We randomly split all 23490 documents into 20000 training and 3490 testing data and employ CCA to learn the projecting vectors of the two original feature spaces. For visual content, we use Dense-Sift feature, color moment and Self-Similarity features to conduct our experiment respectively. We project test data into the subspace using the learned projecting vectors computed during the training procedure. Paragraphs and images are used as queries for text-to-image and image-to-text retrieval, respectively. And the dimensions of the projecting subspace is set as 50, 100, 150, 200, 250 and 300. We calculate the inner product of the query's feature vector and every returned result feature vectors in the subspace. The higher the value of inner product is, the more similar the

[1]Texts in WIKIDATA are descriptors. Texts and category information in NUS-WIDE are noisy tags.

Dataset	Text number	Image number	Other documents	Image feature	Web structure	Category information
Wiki-CMR	74,961	38,804	×	8 types	√	√
Dataset in [2]	1,200	1,200	×	2 types	×	√
Dataset in [4]	2,866	2,866	×	Sift+BOW	×	√
Dataset in [7]	1,987	1,987	×	Sift+BOW	×	√
Dataset in [8]	1,000	300	720	3 types	×	√
WIKIDATA	44,664	237,434	×	×	×	×
NUS-WIDE	5,018-Dim tags	269,648	×	6 types	×	√

Table 1. A comparison of existing cross-modality datasets

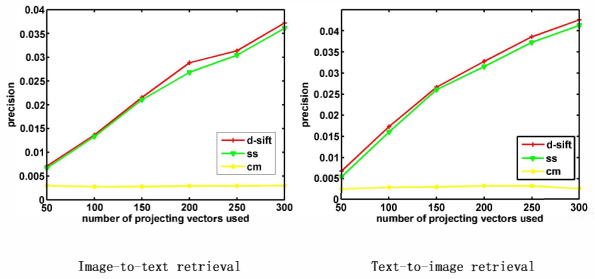

Fig. 4. Top 10 evaluation of CCA

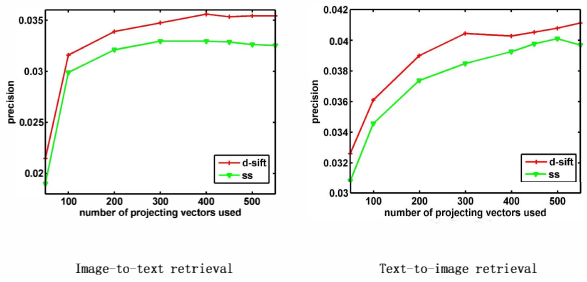

Fig. 5. Top 10 evaluation of KCCA

two documents are. We consider this retrieval process is successful in the top 5 evaluation if the ground-truth is appeared in the first 5 returned documents and successful in the top 10 evaluation if the ground-truth is appeared in the first 10 returned documents.

The result is shown in Fig.4. In both image-to-text and text-to-image retrieval contexts, with higher projected dimension, the performance keeps improving. But the best precision is lower than 4.5% on the dataset. Dense-Sift feature (denoted with d-sift in Fig.4) achieves better result than Self-Similarity feature (denoted with ss in Fig.4) and color moment feature (denoted with cm in Fig.4). For Dense-Sift feature, the result of text-to-image retrieval is better than image-to-text retrieval.

5.2. KCCA

In this section, the kernelized correlation analysis (KCCA) is evaluated. We split the data into two parts, 11745 training and 11745 testing. For visual modality a gaussian kernel (with σ set to be the average distance between images) on the

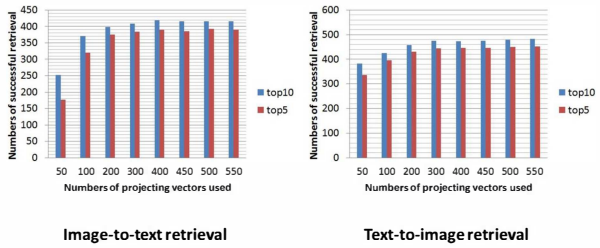

Fig. 6. Comparison of top 5 and top 10 evaluation

Dense-Sift feature or self-similarity feature is used. For textual modality we use the inner product on the reduced 2000-dim features. We project testing data into the projected subspace using the vectors learned during the training. The number of projected subspace is set as 50, 100, 200, 300, 400, 450, 500 and 550. The evaluation criteria is the same as section 5.1.

As can be seen in Fig.5, Dense-Sift feature (denoted as d-sift in Fig.5) outperforms Self-Similarity feature (denoted as ss in Fig.5). When the dimension of projecting subspace is 50, the accuracy of retrieval is poor. When the dimension is 100, we observe great improvement on the precision curve. The performance keeps stable when the number is higher than 400. The best result is about 4%, while result reported in [2] the best result is 90.97%. It indicated that our dataset is more Challenging. Similar with CCA, the accuracy of text-to-image retrieval is better than image-to-text retrieval. As shown in Fig.6, the performance of top 10 accuracy does not outperform much than top 5 accuracy using Dense-Sift feature.

6. DISCUSSION AND FUTURE WORK

In this paper, we introduce a new cross-modality retrieval dataset which will be released in the future and evaluate some cross-modality retrieval algorithms on it. Our future works mainly focus on two parts.

6.1. Increasing the data size

This dataset has already contained more texts and images than the datasets used before. In future work, we will continue to download Wikipedia webpages according to the strategy used in this paper. We also notice that Wikipedia has different

language versions. The same topic item may have different literal descriptions and images in different languages. Such multi-lingual content can not be treated as a simple translation problem since they are produced and edited by different users from different countries. Moreover, we may also consider to collect data from more types of modalities, such as audio and video data in future works.

6.2. Studying cross modality retrieving models

Different from previous datasets used for cross-modality retrieval, our dataset contains more information which better reflects the true Web data structure. In future works, we will conduct evaluation on more state-of-the-art cross-modality retrieval algorithms and show how these approaches take advantage of the rich context information in this dataset. Besides, we will study new cross-modality retrieval algorithms. As shown in the experiments part, existing cross-modality retrieval algorithms do not perform well on this dataset, especially for image-to-text retrieval. A better algorithm which performs well on this dataset using both category and hyperlink information will be the goal of future study.

7. REFERENCES

[1] L. A. Rowe and R. Jain, "Acm sigmm retreat report on future directions in multimedia research," *TOMCCAP*, vol. 1 , no. 1, 2005.

[2] D. R. Hardoon, S. Szedmák, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, 2004.

[3] D. R. Hardoon and J. Shawe-Taylor, "Sparse canonical correlation analysis," *Machine Learning*, vol. 83, no. 3, 2011.

[4] N. Rasiwasia, J. C. Pereira, E. C., G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACM Multimedia*, 2010.

[5] A. Kimura, H. Kameoka, M. Sugiyama, T. Nakano, E. Maeda, H. Sakano, and K. Ishiguro, "Semicca: Efficient semi-supervised learning of canonical correlations," in *ICPR*, 2010.

[6] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *SIGIR*, 2003.

[7] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *ICCV*, 2011.

[8] Y. T. Zhuang, Y. Yang, and F. Wu, "Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval," *Multimedia, IEEE Transactions on*, vol. 10, no. 2, feb. 2008.

[9] J. Tang, R. Hong, S. Yan, T-Seng Chua, Guo-Jun Qi, and R. Jain, "Image annotation by *k*nn-sparse graph-based label propagation over noisily tagged web images," *ACM TIST*, vol. 2, no. 2, 2011.

[10] B. Geng, L. Yang, C. Xu, Xian-Sheng Hua, and S. Li, "The role of attractiveness in web image search," in *ACM Multimedia*, 2011.

[11] T. Kelenda, L.K. Hansen, J. Larsen, and O. Winther, "Independent component analysis for understanding multimedia content," in *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, 2002.

[12] A. Kittur, Ed H. Chi, and Bongwon Suh, "What's in wikipedia?: mapping topics and conflict using socially annotated category structure," in *CHI*, 2009.

[13] D. G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, 1999.

[14] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research*, vol. 11, 2010.

[15] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma, "Image classification by non-negative sparse coding, low-rank and sparse decomposition," in *CVPR*, 2011.

[16] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, 2001.

[17] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, 2002.

[18] Anna Bosch, Andrew Zisserman, and Xavier Muñoz, "Representing shape with a spatial pyramid kernel," in *CIVR*, 2007.

[19] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *CVPR*, 2007.

[20] Jim Giles, "Internet encyclopaedias go head to head," *Nature*, vol. 438, no. 7070, 2005.

[21] Q. Ho, J. Eisenstein, and E. P. Xing, "Document hierarchies from text and links," in *WWW*, 2012.