Context Driven Concept Based Image Retrieval

H. Vahdatnejad, Ph.D. Student
University of Isfahan, I. R. of Iran
Corresponding Author: hamed.vh@gmail.com

M. A. Nematbakhsh, Ph. D.
University of Isfahan, I.R. of Iran
email: nematbakhsh@eng.ui.ac.ir

K. Zamanifar, Ph. D.
University of Isfahan, I. R. of Iran
email: zamanifar@eng.ui.ac.ir

N. Nematbakhsh, Ph. D.
University of Isfahan, I.R. of Iran
email: nemat@eng.ui.ac.ir

Abstract

Several semantic image search schemes have been recently proposed to retrieve images from the web. However, the query context is regularly ignored in these techniques and hence, many of the returned images are not adequately relevant. In this paper, we make use of context to further confine the outcome of the semantic search engines. For this purpose, we propose a hybrid search engine which utilizes concept and context for retrieving precise results. In the proposed model, an ontology is exploited for annotating images and accomplishing search process in the semantic level. Furthermore, the query of the user is modified with the concepts available in the ontology. Next, we make use of search context of the user and augment the query with the information extracted from the user's context to additionally eliminate irrelevant results. Experimental results show that the combination of concept and context is effective in retrieving and presenting the most relevant results to the user.

Keywords: Context, Image Retrieval, Ontology, Semantic Search

Introduction

The growth of digital cameras and even mobile phones leads to the explosion in the number of images and videos available in the web. Therefore, the emergence of an efficient search engine for multimedia data becomes vital. Current web technologies for image search are based on text index. In the text-based search systems, keyword queries are matched with the texts associated to each image such as filename, tag, etc (Yang, Wu, Lee, Lin, Hsu, & Chen, 2008). However, the associated tags typically contain abundant, imprecise, and vague information. In most of the photo-sharing sites, tags and captions are freely entered by the user and contain natural language statements. From April 2007 Exalead and Google have been offering image search which gives better search results but is still based on text attributes (file name, web page, and HTML tag attributes) (Rinc & et.al, 2008). The major problem of the keyword-based image retrieval is low answer quality

(Fan & Li, 2006).The way of applying keyword annotations to images has low ability to investigate semantic relations among keywords such as synonym, homonym and antonym (Hyvonen, Saarela, & Viljanen, 2003). For example, it is impossible to take into account all the synonyms of a keyword in an annotation for every image.

In the past few years exciting progress has been made in advancing the state of the art in the area of multimedia search, but many problems still exist. New ideas and results are based on gathering knowledge from multi-modal content analysis, machine learning, information retrieval, and user interaction (Chang, Ma, & Smeulders, 2007). The main challenge in image search is bridging the semantic gap between bit streams of images and visual content interpretation perceived by humans. Semantic web technologies can be used for solving this problem. Ontology is the foundation of the semantic web, which is used to share a conceptual domain. We exploit the LSCOM (Naphade, Smith, Tesic, & Chang, 2006) ontology to construct a semantic search engine for image retrieval. One of the characteristics of the search engine is that it transforms user query to the concepts defined in the ontology to take into account all of the synonyms. Therefore, it performs concept-based rather than text-based search. As a result, many of the irrelevant outcomes related to homonyms are eliminated, and in addition, related results due to synonyms are included.

Semantic image search outperforms the traditional techniques; however, there are still a lot of situations in which many of the search results are not desirable to the user. In fact, every person issuing a search query on the web accomplishes this with a specific context in mind. Therefore, results lying outside that context are irrelevant to the user (Ramırez & Brena, 2006), though they may be semantically consistent to the user query. Accordingly, we take advantage of the context to further confine the semantic search results and eliminate those locating outside the context. Context is defined in its general form as the information characterizing the situation of an entity, and is composed of time, location, identity, and environment. We utilize the attributes of time, location, and identity (kind) to define the context of images, and the attribute of identity to define the context of users. Therefore, the modified queries of the previous step are augmented by the information available in the context of the user to help us get rid of the results outside the context.

After this introduction, in the next section we review some of the main researches done on the contextual or semantic multimedia search. Then, in section 3, we give a technical presentation of our method. We demonstrate the experimental results in section 4, and finally conclude the paper in section 5.

Review of Literature

In recent years, a number of researches have been concentrated on multimedia search. Unlike the conditions in the 90's, researchers now can easily gain access to millions of images or hundreds of thousands of video shots with some form of annotation either via

formal processes or indirect association (Chang, Ma, & Smeulders, 2007). For example, TRECVID (Smeaton, Kraaij, & Over, 2007) in its 7th year of evaluation supplies hundreds of hours of broadcast news videos from multi-lingual channels. These video programs are segmented into individual shots. Then video shots have been annotated with a number of semantic description labels.

To standardize an appropriate set of sharable concepts, an effort by Naphade, Smith, Tesic, & Chang (2006) has been made to describe a Large-Scale Concept Ontology for Multimedia (LSCOM). In this project about 1000 concepts have been chosen from several categories such as scene, event, object, people, production, and location. Furthermore, a subset of TRECVID videos (more than 80 hours) has been manually annotated by Kennedy & Hauptmann (2006) with 449 of the LSCOM concepts. The result is perhaps the largest video annotation dataset available at the moment.

Labeling process can be done manually or automatically. Often manual labeling processes are exploited in order to acquire complete and accurate annotated results. It is also clear that the quality of annotation significantly affects the quality of search. Although manual annotation yields more reliable results, it is too time consuming. For example, about ten thousand hours of human efforts were consumed in LSCOM to generate about 33 million labels. Each label specifies the presence or absence of a particular concept in a video shot. Moreover, the manual annotation can be done in two ways. In the first approach annotators provide binary-value labels of a concept at a time. In the second approach annotators enter multiple tags related to the image under assessment. The first approach provides much more accurate and consistent annotations than the second one. Afterwards, each video shot or image can be represented by a concept vector as mentioned by Naphade, Basu, Smith, Lin & Tseng (2002) in which each element corresponds to the degree of confidence of a concept.

In the web repositories like web search engines or photo sharing sites, images exist with less reliable labels. The AnnoSearch project that has been accomplished by Wang, Zhang, Jing, & Ma (2006), tends to automatically extend the labels of an image by using free web resources. It performs this by searching the initial keyword of the image known as seed. Then the textual metadata of the result images are clustered in order to find new keywords for the initial image. Furthermore, the new keywords are filtered and irrelevant words are eliminated.

MPEG-7 is a Multimedia Content Description Interface which combines both low level features of image such as color, texture and shape, and high level semantic features. CORTINA (Manjunath & et.al, 2008) and MIRROR (Po & et. al, 2006) use MPEG-7 for accomplishing semantic image search, but both of them are not easily portable. In Rinc & et.al. (2008) a simple approach for semantic annotation and search is presented based on MPEG-7. It uses an open source tool for semantic image annotation. Subsequently, the

annotations are imported into XML database to be available via search with web based semantic image browser. Users can perform semantic image search by using low-level and high-level descriptors. This approach offers a free semantic image annotation and search scheme, which exploits the MPEG-7 standard.

Hammiche, Benbernou, Hacid, & Vakali (2004) extend the MPEG-7 standard by ontology for the domain knowledge representation. Furthermore, they provide some reasoning mechanisms over the MPEG-7 description. In fact, the problem addressed by them can be stated as follows: given a multimedia data query Q and an MPEG-7 database D describing multimedia data extended by an ontology T of a domain, retrieve multimedia data that could semantically answer the client query. For this, the user query and the database description are denoted as *ordered labeled trees*. Hence, the problem of searching multimedia data against MPEG-7 description collections is considered as the problem of a well known *tree embedding*, which is augmented by the subsumption reasoning over the ontology.

In content-based image retrieval (CBIR) techniques such as (Pentland, Picard, & Sclaroff, 1996), (Gupta & Jain, 1997), (Flickner, Sawhney, Niblack, & Ashley, 1995), and (Smith & Chang, 1996), image searching is accomplished based on features automatically extracted from images. SIFT (Lowe, 2004) is a feature generation tool which is invariant to image translation, scaling, rotation, and illumination changes. Ling & Ouyang (2008), propose a two-stage concept searching method for images. In the first stage, an image low-level feature extraction schema is implemented using SIFT. After that, in the second step the latent semantic concept model searching is implemented, and image low level features and global sharable ontology are interconnected. This is done through If-Then association rules, which map the low level features of the image to semantic concepts. In fact, the paper combines ontology and image SIFT features in order for images on the web pages and semantic concepts to be mapped together for semantic searching.

Fan & Li (2006), propose a hybrid model which uses Ontology based reasoning and Bayesian Network model. In addition, the model offers recommendations of search keywords, which are semantically related to the query to assist the users to navigate around the relevant images. In general, the process consists of three steps. In the first step, after providing the search keyword k, the system looks for the target images, which are annotated with the keyword k. In the second step, the ontology reasoner retrieves all the neighboring semantic information of the input keyword. This is performed by retrieving the conceptual graph, finding the matched node representing the concept k, and returning related concepts. In the third step, the Bayesian Network model is used to compute the relevance between the resulted concepts and the initial concept. Afterwards, theses information is used for ranking the results.

Zhuge (2004), proposes a set of primitive semantic links and relevant rules as the basis

of describing and deriving semantic relationships. The semantic links can make a sort of semantic link network, which is the semantic extension of the current hyperlink network. Furthermore, the semantic reasoning and matching used in reasoning and retrieval are accomplished through graph and matrix representation and operations. The approach enables users to acquire the retrieved results as semantic clusters of relevant images rather than a list of isolated images. It also enables users to browse images by traversing along semantic paths with the support of semantic link reasoning.

The IBM multimedia search and retrieval system (Natsev, Tesic, Xie, Yan, & Smith, 2007) consists of two parts: the multimedia analysis engine and the multimedia search engine. The first component exploits machine learning techniques to model semantic concepts in images and videos from automatically extracted visual descriptors. It automatically assigns labels together with associated confidence scores to unseen content. Hence the load of manual annotation is diminished. In the second component, multimedia semantic-based searching is accomplished. Furthermore, it combines content-based, model-based, and text-based retrieval to attain more efficiency in image and video searching.

Jing, Wang, Yao, Deng, & Zhang (2006) implement a semantic image search engine called IGroup, which clusters search results into different groups and enables users to identify their required group at a glance. The advantage of this approach reveals when multiple sub-topics of the given query are mixed together. The IGroup first identifies several semantic clusters related to the query, and changes and modifies the names of the clusters. Afterwards, it exploits the modified names as keywords in the classic search engines like Google. Finally, all the resulted images are assigned to the corresponding clusters.

EXTENT (Tsai, Qamra, Chang, & Wang, 2005) is an image annotation system that combines the context and content information to annotate images. It makes use of the context information of time and location to restrict the image database, and then exploits more expensive algorithms for performing content analysis within the restricted database scope. The authors use a tower dataset and a dataset of images taken from Stanford campus for evaluation of their algorithm. The main contribution of this study is the limited use of context in performing image search.

Yang, Wu, Lee, Lin, Hsu, & Chen (2008) develop ContextSeer which improves search quality and recommends supplementary information (i.e., search-related tags and canonical images) by leveraging the rich context cues, including the visual content, high-level concept scores, time and location metadata. The search engine generally uses the context information in two stages: reranking and suggesting new tags to the user. The main context cues that are utilized in this research are location and time, which are very limited.

In our proposed search scheme, we exploit the semantic search capabilities and enhance it by using context. We make use of context much more than the studies done by Yang,

Wu, Lee, Lin, Hsu, & Chen (2008) and Tsai, Qamra, Chang, & Wang (2005). Furthermore, we utilize the context of the user to direct the search results to her background.

The proposed contextual semantic search scheme

We exploit concept and context for resolving the main problem of image search, which is to decrease the numerous number of images retrieved wrongly for a search query. Furthermore, we intend to retrieve the most relevant images in the first few ones.

Concept is the building block of the semantic web whose aim is to give meaning to the web documents. By making the web documents meaningful, the search scheme changes from text-based to concept-based and hence retrieved images become more accurate and many of the irrelevant results vanish.

Another significant attribute of queries and images is their diversity in context; For example, images with the same entity may be completely different as they represent dissimilar features of it. For better explanation, consider three images of an animal such as a tiger. The first one is a beautiful, high resolution, and natural image, which may be used as a wallpaper or background image in the desktop of a computer. The second one is an art painting, which has been sketched by a famous artist and may be utilized as a tableau. Finally, the last one is an anatomical or physiological image, which may be utilized as a scientific image. On the other hand, different people like a computer science student, an artist or a biologist have different aims in searching the keyword "tiger". In other words, each one probably searches the keyword "tiger" in his own context. The intention of the computer science student or artist may rarely be a physiological or biological image of a tiger. In general, each individual usually issues a query in his context (i.e. his identity). Therefore, if we have modeled the context beforehand, we can confine the search results and retrieve images related to or exactly in the context of the user. In the following subsection, we discuss the use of concept and context in the image search with more details.

Conceptual search

We take advantage of the semantic web technologies in the proposed search engine. The main challenges in the semantic search are constructing a shared ontology and annotating images by using the concepts defined in it. We make use of the LSCOM developed by Naphade, Smith, Tesic, & Chang (2006) as the base ontology, and moreover extend it with new concepts. Afterward, the images are annotated manually with the concepts available in the ontology. The annotation is accomplished by constructing a concept vector of zeros and ones for each image. Each element of the vector corresponds to a concept. If an image represents a special concept, the corresponding element in the concept vector is one, otherwise it is zero. The major advantage of manual annotation is its high precision comparing to the automated tools. When a user issues a query, the query is

mapped to the concepts of the ontology and the selected concepts are treated as new keywords and searched against the metadata associated to each image. In other words, we modify the keyword using ontology and utilize the changed keyword in the subsequent process.

Contextual search

We further devote to each image the attributes of time, location, and identity. Most of digital cameras provide facilities such as clock and GPS for showing and recording time and location. These utilities can be exploited to find the time and location metadata of the context. The identity of an image represents kind and application of the image (e.g. scientific, art, news, family, etc). On the other hand, the context of each user is extracted and utilized to confine the retrieved results to the desirable images (i.e. the images within their context). We exploit the identity of a user as his context. To extract identity of a user, the information available in their profile as well as web pages which have been visited recently are utilized. Afterward, the extracted metadata is mapped to the concepts of the ontology and the context of the user is represented as a concept vector in the ontology.

After issuing a query by the user, throughout the previous stage the system converts it to the concepts available in the ontology. Then the query is augmented by the information available in the context (identity) of the user. For this, the mean similarity between each concept in the user's context and the concepts available in the modified query is computed, and then the concept in the user's context which has the highest similarity is selected and added to the query. In other words, the semantic similarities between each concept in the user's context and all of the concepts in the query are computed and their average is calculated. Hence, the concept in the user's context which has the maximum mean similarity is selected and utilized for expanding the query. Afterward, the augmented query is utilized as the new keyword in the search engine.

Experiments

When a user is searching through entering a keyword query into a search engine, he has a distinct context and a specific goal in mind. The users' queries are usually vague and contain between one to three keywords. The outcome of the search engine may satisfy the search criteria but it frequently fails to meet the user's search objective.

As there are no public benchmarks for the shared user photo collections, we resort to building up the dataset by ourselves. We constructed a dataset involving 5000 images. The images are mostly indicating places and computer related objects. We extend the LSCOM ontology and annotate the images of the dataset using the available concepts.

Furthermore, for each image, any of the time, location and identity metadata that is applicable is appended. The time metadata involves the date and time information and can

be exploited when queries like Christmas tree, Oktoberfest, or morning moon are issued. Location metadata can be extracted from location aware cameras or mobile phones. Therefore, it is mapped to the name of a place such as name of a street, city, or country or even a well-known institute or organization such as laboratory, museum, etc. The location metadata can help in many situations; for example, when the query is "Eiffel Tower" and it is a concern that similar towers in another countries or even drawings of it may be retrieved, the location metadata of images can be utilized to eliminate irrelevant results. Finally, the aim of the identity metadata for an image is to specify kind and application of it. This attribute is expressed as a concept vector consisting of concepts representing kind and application of the image. For example, an image of a tower may be of the kind of engineering, in which it illustrates the structure and architecture of the construction, or of the kind of perspective in which it shows beauty and attraction. Consequently, the latter may be utilized for introduction and advertisement of the place to attract tourists for visiting. As another example consider a physiological image and an art painting picture of an animal. The former is of scientific, biological, and physiological kind but the latter is of kind of art and painting.

On the other hand, the context of each user is extracted as his identity, and is expressed as a concept vector. This concept vector represents the user's personality, interests, profession, etc; for instance, the context of a computer science student may compose of engineering, computer science, research, guitar (as he plays guitar), music, and football.

To denote the role and significance of context in the search scheme, we compare the proposed conceptual and contextual approach to the semantic search scheme, which utilizes the concept and ontology but ignores the context in accomplishing search. Two measures are typically exploited to evaluate the information retrieval applications: precision and recall (Clarke & Willett, 1997). At the beginning of our experiment, we assumed that all the images are well annotated. Therefore, our experiment result is not influenced by this factor. Under this assumption, most of the relevant results are retrieved, and hence recall is not an important criterion for evaluation. However, precision is an important parameter which is defined as the number of positive retrieved results to the total number of retrieved results.

We conducted our experiment in two phases. First we constructed keyword queries from the concept terms available in the ontology. For this, we take some annotation words randomly from the original annotations and make use of them as simulated queries. The queries used are between one to three keywords. Usually, as the number of keywords in a query increases, the search query becomes clearer and less ambiguous. In Figure 1 the results of comparison are illustrated for all generated queries. The evaluation results verify that the proposed approach does better than the conceptual model in increasing the

precision metric. In particular, when user query contains time or location terms, our scheme yields better outcome. Moreover, the proposed approach performs excellent in making use of context to disambiguate vague queries.
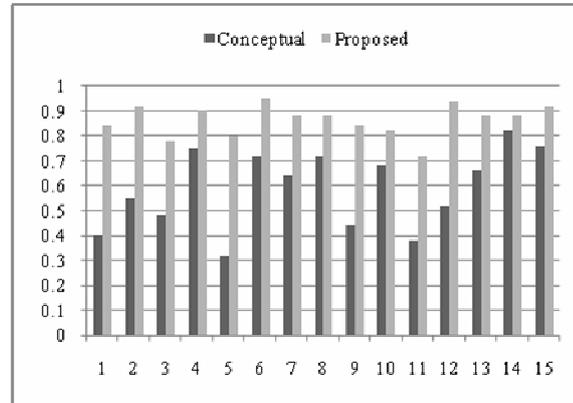

Figure 1. Precision results of the conceptual and the proposed scheme for 15 randomized generated queries.

In the second phase, we take into account some special or ambiguous queries to simulate the real situation. Table I shows the comparing results in terms of precision. In the first query, the user intends to retrieve images of the Milad tower (in IRAN), which were taken in 1998. The construction of the Milad tower was not finished in that year; hence, the user probably means images illustrating the level of construction of the tower in that time. For this query, the traditional and also semantic schemes exhibit a very poor performance. However, the proposed approach makes use of the time attribute (context) of the images and therefore, achieves a very good performance. In the second query, the user has in mind to see images of the Incan pyramids, located in the South America. Similarly, the proposed approach utilizes the location component of the images to retrieve relevant results, and as a result, performs much better than the conceptual model. The third query is issued by a computer science student, who would like to retrieve images illustrating the ant colony algorithm. As can be seen, the proposed scheme does well in exploiting user context to guess the purpose of the query. It achieves this by augmenting the user query by the user context (computer science). The last query is issued by a history researcher, who intends to retrieve images of the Incan pyramid. However, the conceptual search scheme retrieves many irrelevant images such as a street or a computer game with the name Incan pyramid. On the other hand, the proposed model makes use of the user context (history) to augment the query and retrieve much more relevant results.

Table 1

Precision results of four special queries

Query	User context	Conceptual	Proposed
Milad tower	Not specified	0.07	0.90
Pyramid South America	Not specified	0.47	0.87
Ant colony	Computer science	0.03	0.70
Incan pyramid	History	0.67	0.93

Conclusion

In this paper, we propose and evaluate a new hybrid model for image search. The model makes use of the semantic web technologies to accomplish the search operation in the semantic world. It further makes use of the context to reduce the results locating beyond the search context. For this purpose, it takes advantage of the image and user context. As a result, the proposed search engine performs well in retrieving images especially for queries that contain contextual data. Examples of this contextual data are time and location. We demonstrate through experimental results that making use of user context can significantly increase the accuracy of a search engine. In addition, we verify that the proposed approach does well in disambiguating a query using user's context.

One of the proposing directions for future work is to implement the proposed search system over the Internet and make use of the diverse images available on it. For this purpose, we intend to parallelize the search engine to keep its efficiency over large databases. Furthermore, we plan to include video data in the search dataset and extend the system to perform multimedia search through the Internet.

References

Chang, S., Ma, W. & Smeulders, A. (2007). Recent advances and challenges of semantic image/video search. *2008Iinternational Conference on Acoustics, Speech, and Signal Processing.* Hawaii, USA: IEEE Computer Society.

Clarke, S. J. & Willett, P. (1997). Estimating the recall performance of search engines. *Aslib Proceedings, 49* (7), 184 - 189.

Fan, L. & Li, B. (2006). A hybrid model of image retrieval based on ontology technology and probabilistic ranking. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 477-480. Washington, USA: IEEE Computer Society.

Flickner, M., Sawhney, H., Niblack, W. & Ashley, J. (1995). Query by image and video content: The QBIC system. *IEEE Computer , 28* (9), 23 - 32.

Gupta, A. & Jain, R. (1997). Visual information retrieval. *Communications of the ACM, 40*, 70-79.

Hammiche, S., Benbernou, S., Hacid, M. & Vakali. (2004). Semantic retrieval of multimedia data. *Proceedings of the 2nd ACM International Workshop on Multimedia Databases*, 36-44. Washington, DC, USA: ACM.

Hyvonen, E., Saarela, S. & Viljanen, K. (2003). Intelligent image retrieval and browsing using semantic web techniques: A case study. *The International SEPIA Conference.*

Jing, F., Wang, C., Yao, Y., Deng, K. & Zhang, L. (2006). IGroup: A web image search engine with semantic clustering of search results. *Proceedings of the 14th Annual ACM International Conference on Multimedia*, 497-498. Santa Barbara, CA, USA: ACM.

Kennedy, L. & Hauptmann, A. (2006). LSCOM lexicon definitions and annotations version 1.0. *DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia.* Columbia University.

Ling, Y. & Ouyang, Y. (2008). Image semantic information retrieval based on parallel computing. *In Proceedings of the ISECS International Colloquium on Computing, Communication, Control, and Management* , 255-259. Guangzhou: IEEE Computer Society.

Lowe, D. (2004). Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision , 60* (2), 91–110.

Manjunath, B. S. & et.al. (2008). *CORTINA image search.* (University of California) Retrieved 2008, from http://cortina.ece.ucsb.edu

Naphade, M. R., Basu, S., Smith, J. R., Lin, C. Y. & Tseng, B. (2002). Modeling semantic concepts to support query by keywords in video. *In Proceedings of the IEEE International Conference on Image Processing. 1*, 145-148. IEEE Computer Society.

Naphade, M., Smith, J. R., Tesic, J. & Chang, S. F. (2006). Large-scale concept ontology for multimedia. *IEEE MultiMedia, 13* (3), 86-91.

Natsev, A., Tesic, J., Xie, L., Yan, R. & Smith. (2007). IBM multimedia search and retrieval system. *6th ACM International Conference on Image and Video Retrieval.* ACM.

Pentland, A., Picard, R. W. & Sclaroff, S. (1996). Photobook: Tools for content-based manipulation of image databases. *International Journal of Computer Vision*, 233-254.

Po, L. & al. e. (2006). *MIRROR.* (Department of Electronic Engineering, City University of Hong Kong, China) Retrieved 2008, from http://abacus.ee.cityu.edu.hk/~mpeg7.

Ramırez, E. H. & Brena, R. F. (2006). Semantic contexts in the internet. *Fourth Latin American Web Congress*, 74-81. IEEE Computer Society.

Rinc, S. & et.al. (2008). Towards standard semantic image annotation and search. *International Workshop on Content-Based Multimedia Indexing,* 484-488. London: IEEE Computer Society.

Smeaton, A., Kraaij, & Over, P. (2007). TREC video retrieval evaluatio*n*. Retrieved from http://wwwnlpir.nist.gov/projects/trecvid/

Smith, J. R. & Chang, S. F. (1996). VisualSEEk: A fully automated content-based image query system. *ACM Multimedia,* 87-98.

Tsai, C., Qamra, A., Chang, E. Y. & Wang, Y. F. (2005). EXTENT: Inferring Image Metadata from Context and Content. *IEEE International Conference on Multimedia and Expo* (pp. 1270-1273). Amsterdam: IEEE Computer Society.

Wang, X. J., Zhang, L., Jing, F. & Ma, W. Y. (2006). AnnoSearch: Image auto-annotation by Search. *IEEE CVPR.* New York: IEEE computer society.

Yang, Y., Wu, P. T., Lee, C. W., Lin, K. H., Hsu, W. H. & Chen, H. (2008). ContextSeer: Context search and recommendation at query time for shared consumer photos. *ACM Multimedia* , 199-208.

Zhuge, H. (2004). Retrieve images by understanding semantic links and clustering image fragments. *Journal of Systems and Software, 73* (3), 455-466.