Lau, C. and Tjondronegoro, Dian W. and Zhang, Jinglan and Geva, Shlomo and Liu, Y. (2007) Fusing Visual and Textual Retrieval Techniques to Effectively Search Large Collections of Wikipedia Images. In Proceedings 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006 4518/2007, pages pp. 345-357, Dagstuhl Castle, Germany.

Fusing Visual and Textual Retrieval Techniques to Effectively Search Large Collections of Wikipedia Images

C.Lau, D.Tjondronegoro, J.Zhang, S.Geva, Y.Liu

Faculty of Information Technology, Queensland University of Technology,
2 George Street, GPO Box 2434, Brisbane, QLD 4001 Australia
{andy.lau, dian, jinglan.zhang, s.geva}@qut.edu.au
y53.liu@student.qut.edu.au

Abstract. This paper presents an experimental study that examines the performance of various combination techniques for content-based image retrieval using a fusion of visual and textual search results. The evaluation is comprehensively benchmarked using more than 160,000 samples from INEX-MM2006 images dataset and the corresponding XML documents. For visual search, we have successfully combined Hough transform, Object's color histogram, and Texture (H.O.T). For comparison purposes, we used the provided UvA features. Based on the evaluation, our submissions show that Uva+Text combination performs most effectively, but it is closely followed by our H.O.T- (visual only) feature. Moreover, H.O.T+Text performance is still better than UvA (visual) only. These findings show that the combination of effective text and visual search results can improve the overall performance of CBIR in Wikipedia collections which contain a heterogeneous (i.e. wide) range of genres and topics.

1 Introduction

Recently there has been a renewed spurt of the research activity in multimedia information retrieval [1-3]. This can be partly due to the rapid proliferation of the Internet. The current World Wide Web increasingly uses structured XML documents that contain not only the text information, but also other media information such as images, videos, and speech. Images have been playing an important role in human life as one of the most important information resources. The amount of images on the Web is increasing exponentially due to the advent of various digital devices such as digital cameras and scanners. Nowadays the problem is not *finding* information, but how to find *useful* information efficiently and effectively. Image retrieval techniques have attracted intensive interests from both the industry and the research communities.

Currently there are two paradigms in image retrieval: text-based and content-based. Text-based image retrieval techniques perform retrieval using keywords. Most of the state-of-the-art systems, such as Google and Yahoo!, belong to this paradigm. Although text description is closer to the concepts in human mind, they rely heavily on manual annotation of images for keyword matching, which is tedious and costly. In addition, it is hard to define a unified set of keywords to be used for annotation. Although the concept of using game players to help label images on the web via an online game is very interesting [4], the final effect of this approach to effectively label images still needs to be verified due to Internet abuse [24]. Content-based image retrieval turns to visual features for searching similar images, which alleviates the burden of manually labelling images. However, due to the well-known semantic gap between low-level visual features and the corresponding high-level concepts, visual features are generally not sufficient for searching similar images [5].

In order to mitigate this semantic gap, multi-modal image retrieval, which uses both text and content-based searching, is attracting uprising interests [6]. It has been proved that better retrieval results will be achieved by appropriately fusing different modal information [7]. In most cases, especially on the Web, images do not exist separately. Instead, there is much relevant information surrounding these images. Combing different modality, e.g. images and textual information, would improve the retrieval accuracy; fusion of the image and the text will also make the query more flexible for users. A user could search the image database by the image and/or the text. Although combining the image with the text together has been studied just recently, there are still some open issues needing further study, such as how to combine the content and text-based image retrieval results together.

The main goal of this research is to develop and evaluate algorithms for structured document retrieval systems using comprehensive database of XML documents containing text and image information. The document collection used is provided by the INEX 2006 organizing committee. The corpus contains 166,559 images in formats such as PNG, JPG and GIF. This complex selection of images depicting both natural and man-made objects (such as landscape, people, animals, buildings, and logos) comes in different sizes as well as different color depths. This project aims at creating a Content Based Image Retrieval (CBIR) system which can deal with a large set of heterogeneous images and will work together with an existing text-based search engine in order to elevate the quality of the final retrieval results.

In this research, a text and an image-based search engine are used concurrently and the results are obtained by fusing two or more independent result sets. Therefore, the primary contribution is to answer the challenging question of how to fuse multiple results to form optimal query results for users. This involves the fusion of the XML document retrieval results of multiple search algorithms on the same collection. Queries consist of both text (keywords) and/or example images. Two document retrieval algorithms are used: a text-based retrieval using a *TF-IDF* variant, and an image-based retrieval using image feature similarity.

2 Related Work

This section summarizes some previous work in image feature extraction and multi-modal image retrieval.

Image feature extraction

Compared with text-based image retrieval which takes advantage of keywords or metadata such as captions, authors, and textual descriptions, content-based image retrieval is more complicated and has to extract the appropriate visual features first [5].

Color is one of the most widely used visual features. The choice of a color space is of great importance to the proper color-based image retrieval. HSV and YCbCr are two commonly used color spaces which can better model the human color perception [8]. The histogram, which describes the distribution of colors in an image, is a traditional representation of the "color" feature. It is usually high dimensional and contains more global information of the image.

Texture is another important feature which represents some important aspects of many real-world images, such as bricks, coins, trees, etc. Texture has characteristics such as periodicity and scale, and could be represented in terms of direction, coarseness, contrast, etc. In this sense, texture features contain the high-level semantics for image retrieval [5, 8]. Texture features could be divided into two categories: the structural texture and the statistical texture. The structural method represents the texture by identifying structural primitives and their location rules, which consists of morphological operator and adjacency graph. The statistical approach, which is one of the earliest methods to classify textures, describes texture by the spatial distribution of image density.

Hough transform is a kind of feature extraction method which can identify objects with a particular shape in an image. It includes two kinds of transform methods: the classical transform and the generalized transform. The classical Hough transform is usually used to detect the regular curves such as lines, circles, ellipses, etc. The generalized transform is applicable for the detection of positions of arbitrary shapes which cannot be described by using the simple features.

We will use color histogram, statistical texture, and generalized Hough transform in our experiment.

Fusing Image and Text retrieval

One challenge for image retrieval is to find a simple but effective way to form a query. Most content-based image retrieval systems support query-by-example in which users should provide visual example of the contents they seek. In this case images are searched on the basis of matching of content features such as color, shape, and texture. Therefore, this method is more intuitive, and an appropriate key-image is dispensable to start a query. However, this query method has two drawbacks. Firstly, a user may not find such an appropriate image which can represent the use's query need completely in some cases. Secondly, the representation of the image is not as flexible as the textual description, and most users have been used to adopting keywords to start their query and describe their needs.

Text-based image retrieval can address these problems, which is based on the assumption that the textual description can express the semantics of images [9]. It allows users to search images by specifying their own query in terms of a limited vocabulary of semantic concepts. But *synonymy* and *polysemy* are two large existing problems in information retrieval [10]. Many words have more than one distinct meaning. Users in different contexts may express the same needs by using different terms. In addition, there are many ways to describe the same object. In different contexts or when used by different people, the same terms can be taken differently. The prevalence of *synonymy* and *polysemy* tends to degrade *precision* and *recall* performance of the system.

It is believed by many researchers that combining the keyword-based approach and the content-based approach together can benefit from the strengths of both paradigms, and these two paradigms can supplement each other. R. Zhang et al. [6] in their paper show that multi-modal image retrieval is a promising way to improve image retrieval and enhance users' querying modalities. There are several ways for multi-modal fusion, such as linear combination, min-max aggregation, and voting production combination [7]. These methods can be classified into two categories: fusion at feature level and fusion at output level. It has proven that the fusion on output level outperforms the fusion on feature level in most cases [7].

E. Chang et al. [11] suggest that the user can start a query by a few keywords, and after a few relevant images are returned, the image features with their annotation can be used to perform a multi-modal query refinement. J.L. Martinez-Fernandez et al. presents similar ideas [12]. They refine the text-based search results with the addition of content-based image retrieval. Based on their successful previous work on organizing images according to the visual similarity for image retrieval, D. Cai et al. [13] use low-level features to cluster images into semantic clusters obtained by the textual features. R. Basancon et al. [14] presents the opposite idea. They first search the candidate images using content-based methods, and then use the textual description of these candidate images as query keywords to search again.

However, regardless of whether the text-based results are refined using content-based methods or vice versa, it is still insufficient to improve the performance of image retrieval. The reason is that owing to the intrinsic drawbacks of text- and content-based approaches, it is hard to confirm which method could achieve better results in terms of a specific query. It is our hypothesis that the late combination would assure better searching results. D. Tjondronegoro et al. [15]indicate, by a series of experiments, that the results of content-based retrieval with the help of text-based retrieval is much better than any individual text-based or image-based retrieval. During the work in the INEX 2005, they designed two search engines: an XML document search engine using both structural and content information and a content-based image search engine. When a query is submitted, these two engines work respectively at the same time, and then the retrieval results are merged together by treating the visual features as the text terms. The question of how to fuse content-based retrieval with text-based retrieval, however, still needs further consideration. There are no common agreements on the fusion approaches.

3 System Architecture

The framework of the prototype system is shown Figure 1. Users interact with a Graphical User Interface (GUI) to write queries and browse results. The GUI merges results from an integrated document searching that fuses image-text results. The database stores images and the XML documents which display (i.e. provide links) them. Three (3) image related tables are used in the database: Images, Features, and AdhocXML (as shown in Figure 2). The *Images* table stores the information about the images, including the image ID (e.g. 22625), its original filename (e.g. *DVD-RW_Spindle.jpg*), the collection set and the subfolder name. The *Features* table stores the image feature extracted using the feature extractor. The *AdhocXML* table stores the Adhoc XML filename and the absolute XPath of where the related image appears in the document. This allows the system to back-track the original XML document which uses the image and also allows it to fuse the image and text search results. The selected image features for this experiment include color histograms (RGB/HSV/YCbCr), textures, detectable lines (Hough transformation), and the UvA features provided by INEX2006 organizing committee (developed by a research group in University of Van Amsterdam). The UvA feature uses Natural Image Statistics, Color invariant edge detection and regional descriptors to represent an image. These features of every image are stored in the database for CBIR. All image features are represented by vectors. Different distance measures such as the correlation coefficient, Euclidean distance, and Manhattan distance have been implemented and can be chosen from the Graphical User Interface. The Euclidean distance between two vectors is used for the final submission. All individual distances are normalized to the range of [0, 1].

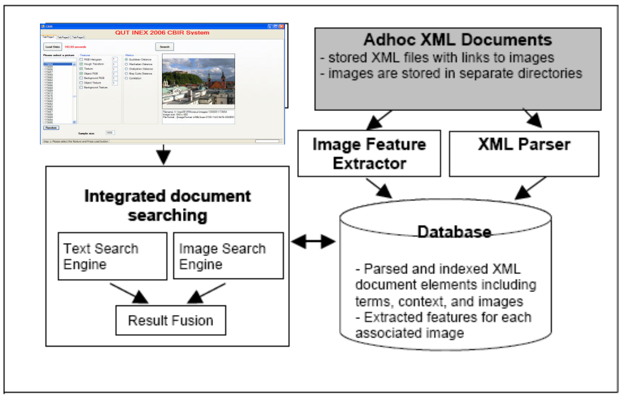

Figure 1: System Architecture

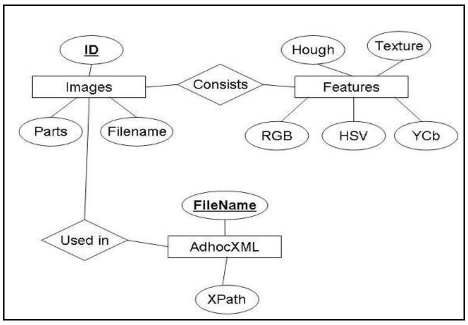

Figure 2: Database Entity-Relationship

Graphical User Interface (GUI)

Using the GUI, users can choose which features to be combined and adjust the weights of each individual feature, as well as the metrics of distance calculation. The GUI also provides a preview picture box and the information of the image.

Memory Management

Due to the large amount of data and high dimensionality of the image feature, a data access technique is required to avoid system resources overload (while calculations are performed in the memory). The system is designed to perform calculation in sequential order. Each time a block of 100 image features is read and stored in the memory buffer, only the normalized distance (from the query image) is kept in the buffer, while the rest are overwritten. This method is found to be efficient and reliable (i.e. no buffer overflow) during our experiments.

4 Using Visual Retrieval Techniques

The image features used in this project are color histogram, texture and detectable lines (Hough transform).

Color histogram

Color histogram has been widely used in CBIR to represent the global image color attribute. For our experiment, we compared the use of RGB, HSV, and YCbCr color spaces. RGB color histogram is invariant with translation and viewing-axis rotation, and varies only slowly with the angle of view. Therefore, it is particularly suited for recognizing an object of unknown position and rotation within a scene. Moreover, translation of an RGB image into the illumination invariant rg-chromaticity space allows the histogram to operate well in varying light levels [16]. HSV Histogram is one of the most commonly used by people (especially artist) to select color. It is a cylindrical coordinate system where the angle defines hue, the radius defines saturation, and the vertical axis defines color of an image. As compared to RGB, which is an additive model, HSV encapsulates information about each color in a more understandable and recognizable manner to humans who perceive color via its color name, vibrancy and darkness/brightness. Similarity distance between HSV color space often performs better than RGB (e.g. [17]). The YCbCr color histogram represents the luminance level using a single component, Y, and other color information is stored individually using Cb and Cr. This feature is frequently used in skin color detection projects (e.g. [18]) as values extracted from Cb and Cr can effectively represent skin color of most human races. Figure 3 illustrates that these various color spaces will return different results for the same query, thus we aim to compare the effectiveness of each color representation for different topics.

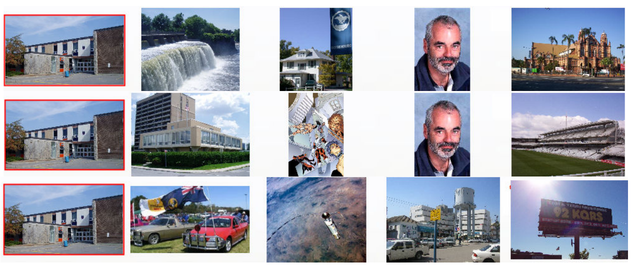

Figure 3: Sample Results from Retrieval using Various Color Histogram (From top to bottom: RGB, HSBV, YCbCr Results)

Object Histogram

A pre-processing of object extraction is performed on every image and hence only the histogram of the extracted segments is calculated. This technique is useful when performing queries with standout and dominant object(s). The algorithm described in [19] can efficiently extract building and human apart from the background. The object extraction accuracy generally increases when the color of the background is much different from the object itself. Several researches, including our previous work [15, 20], shows that object extraction can be used to enhance the image retrieval system performance. Figure 4 shows an example of object-extraction result on an airplane image which has a distinctive object features.

Hough Transform

Hough transform is widely used in digital image processing to detect arbitrary shapes and identify lines (e.g. [21]). For the Wikipedia images dataset, we expect that this method can effectively distinguish some topics such as buildings, coins and some arbitrary shapes since they have unique lines strength characteristics. Figure 5 shows the use of Hough transform to detect coin shapes.

Figure 4: Object extraction algorithm

Figure 5: Hough Transform detect Coin Shapes

Texture

Texture can often represent the visual properties of an image which are unable to represent by color intensity alone. For example, while sea and sky are both blue, but their texture features are different In this project we used a 6D texture-vector which comprises of the intensity, contrast, smoothness, skewness, uniformity and randomness, using a method which has been described in our previous work [19]. We expect that texture feature is likely to perform well in locating images with distinctive and strong unique patterns.

UvA Features

This feature uses Natural Image Statistics, Color invariant edge detection and regional descriptors to represent an image. The paper [22] presents a 120D features using different texture parameters. In our project, we will compare UvA features with our features set.

Similarity measurement metrics

We implemented and compared the performance of the following similarity measurement metrics: *Euclidean* [23], *Manhattan* [23], *Chebyshev*, *Bray Curtis* and *Correlation*. Euclidean distance is commonly used in most of the image retrieval system and Correlation is able to indicate the strength and direction of 2 vectors to show how closely one image correlate to another. To represent the greatest of their differences along any coordinate dimension, the Chebyshev distance of any two points is defined as:

$$D = \max(|x_2 - x_1|, |y_2 - y_1|)$$

Whereas, Bray Curtis distance is defined as:

$$D = \frac{\sum_{k=1}^{n} |x_{ik} - x_{jk}|}{\sum_{k=1}^{n} (x_{ik} + x_{jk})}$$

5 Using Text Retrieval Techniques

The system that we have used for XML text based retrieval is the GPX search engine [23]. The system is described in detail in this proceedings collection under the title "GPX - Gardens Point XML IR at INEX 2006". The MM Fragments task was applied without any changes to the system. For the MM Images task we had to index the new collection but no other work or modifications were required. The system ignored search by image source. All clauses such as about (.,src:988336) were simply ignored by the system. It was left to the image retrieval system to take account of example images and fuse the CBIR results with the text retrieval results, as outlined in the next section.

6 Combining Visual and Textual Search Results

The fusion of search results is performed in twofold: 1) combination of image features, and 2) fusion of text-based search and image-based search results. To combine multiple image features, a weighted sum is performed on the individual distances calculated using different features. This is based on the hypothesis that the combination of more evidence will increase the reliability of the similarity calculation.

To combine the results of text-based and image-based search engines, we use the result of the text search as the base, while the image search result is used to boost the confidence. Our text-image fusion algorithm relies on a hypothesis that if a document appears in both text and image based search results, then the final ranking of the document should be pushed to the top. This hypothesis has not been thoroughly proven as our experiment shows that sometimes the individual search results (either image-based or text-based) can still be better than the combined. More details will be discussed in Section 7 (Evaluation).

7 Evaluation

Our system's prototype is designed using VB.Net language in Microsoft .NET 2.0. The database used is Microsoft Access. The image feature extraction is performed using MATLAB with Image Processing Toolbox. The image collection used is provided by INEX 2006 (http://inex.is.informatik.uni-duisburg.de/2006/). The corpus contains 166,559 images of different format such as PNG, JPG and GIF. All of the images come in different size, different color depth and present various type of content from landscape, buildings, people, and buildings. Table 1 summarizes the topics of INEX 2006 MM tasks. For each task, we will re-interpret the queries as "Find *other* images about X like this one".

Topic ID	Topic Title
1	Sunset sky
2	Historic castle
3	Barcelona
4	Bactrian Coins
5	Masjid (Mosque) Malaysia
6	Da Vinci's Color Paintings
7	Mountain for hiking
8	Images of bees with flowers
9	Golden gate bridge
10	Stringed musical instruments
11	Frank Lloyd Wright Buildings
12	Rodin "Burghers of Calais"
13	Official logos of parties in Sweden

Table 1: INEX 2006 Image Retrieval Topics

We experimented with various combinations of features and found that the combination of Hough transform (H), the Object color histogram (O) and the Texture (T) with equal weighting performs best (among other combinations).

To illustrate the benefits/weaknesses of using text- or image- results only, or using the fusion of text-image results, Figure 6 and 7 shows the text-only search results and fused search results of Topic 9, respectively. For this case, our system is able to refine the text search by bringing forward the 2 visually similar images with lower ranking.

To illustrate the benefits of our H.O.T, Figure 8 and 9 shows the results for Topic 4 using visual features only and fusion of text-image, respectively. In this case, visual features effectively locate images with similar color, shapes and texture. However, with the fusion of the text and image, the results are actually affected by the *Bactrian camel* which is irrelevant.

After evaluating various similarity metrics, we found that the results of applying different similarity measurement metrics are similar and we decided to use *Euclidean* distance as it is one of the most commonly used similarity measurements,

Figure 6: Text Results of Topic 9

Figure 7: Fused (i.e Text-Image) Search Results of Topic 9 (This shows that fusion of text and image search results is better than text search alone)

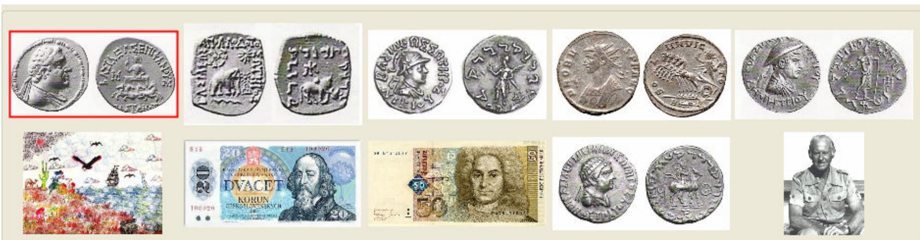

Figure 8: H.O.T image-based search result for Topic 4

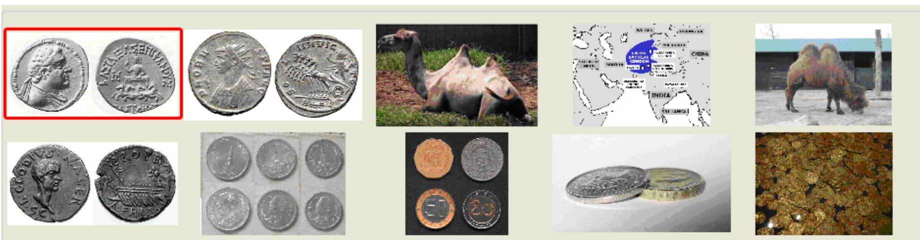

Figure 9: Fused search results of Topic 4 (This shows that H.O.T feature performs better when it is used without merging with text results)

Figure 10 depicts the official evaluation result from INEX 2006 for QUT submissions (QUT's performance is highlighted- while the non-highlighted are from other participants). Based on the interpolated precision-recall performance, the best technique is UvA+Text results. However, the H.O.T (visual only) feature performed at a similar quality, and in fact is better than UvA (visual) alone. This shows that for certain image retrieval task, a combination of simpler low-level features can be sufficient to produce a decent result. However, visual features cannot fully represent the semantics for an image, which is why text search needs to be exploited to improve the accuracy of search results. Moreover, XML documents which use images usually provide useful contextual information which sometimes can be close to users' search intention. Furthermore, it is worth noting that detectable lines and shape description are effective to search images with distinctive objects such as cars, airplanes, coins, and buildings.

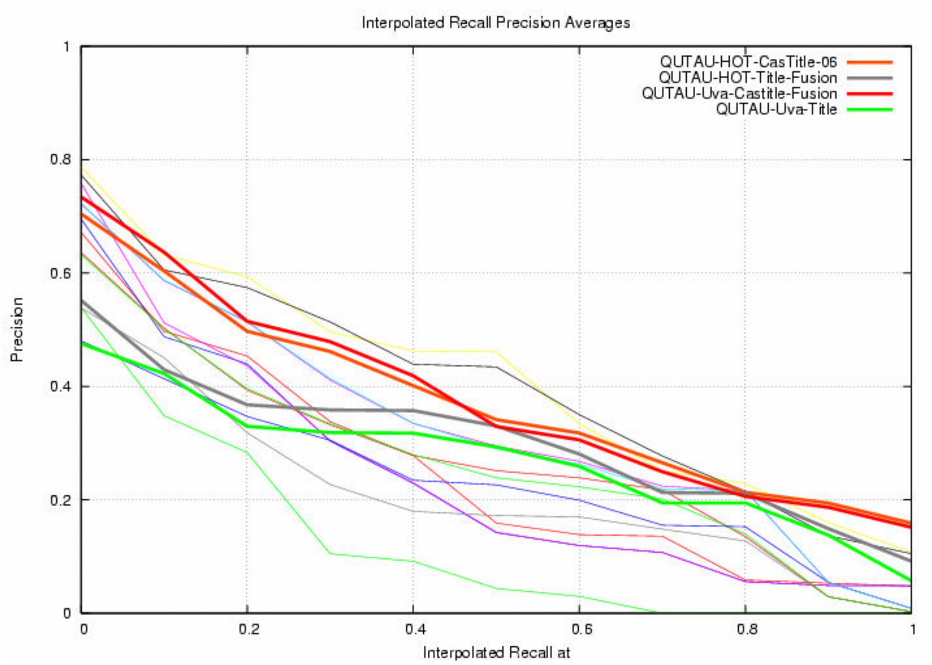

Figure 10: Graph of Interpolated recall precision averages

8 Conclusions and Future Work

The experiment result shows that once the text-based image searching results are refined using content-based image searching results, the final combined result is generally better than the individual text or image-based searching results.

The semantic relation between keywords has not been investigated, and remains a subject of future investigation as it is one of the main reasons incurring low precision in image retrieval. Ontology contains concepts and their relations, therefore introducing ontology into multi-modal image retrieval will not only help better analyze the text, but are also useful for keyword matching. In this sense, text-based retrieval will become concept-based retrieval. We hypothesize that the fusion of concept and content will achieve much better results than the fusion of text and content.

In addition, fast access of the image database needs to be studied as it is very large and image features are usually high dimensional (from 100 to 1000) and slow down content-based matching.

References

[1] Michael S. Lew, Nicu Sebe, Chabnane Djeraba, and R. Jain, "Content-based multimedia information retrieval: state of the art and challenges," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 2, pp. 1-19, 2006.

[2] M. L.Kherfi, D. Ziou, and A. Bernardi, "Image retrieval from the World Wide Web: Issues, techniques and systems," *ACM Computing Surveys*, vol. 36, pp. 35-67, 2004.

[3] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1349-1380, 2000.

[4] L. von Ahn and L. Dabbish, "Labeling images with a computer game," *Proceedings of the 2004 conference on Human factors in computing systems*, pp. 319-326, 2004.

[5] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, pp. 262-282, 2007.

[6] R. Zhang, Z. M. Zhang, M. Li, W.-Y. Ma, and H.-J. Zhang, "A probabilistic semantic model for image annotation and multi-modal image retrieval," *Multimedia Systems*, vol. 12, pp. 27-33, 2006.

[7] Hanghang Tong, Jinrui He, Mingjing Li, and C. Zhang, "Graph based multi-modality learning," presented at Proceedings of the 13th annual ACM international conference on Multimedia, Hilton, Singapore, 2005.

[8] T. Gevers and A. W. M. Smeulders, "Content-based image retrieval: an overview," in *Emerging Topics in Computer Vision*, S. B. k. Gerard Mediomi, Ed. USA: IMSC, 2004, pp. 333-384.

[9] Z. Gong, Q. Liu, and J. Zhang, "Web image retrieval refinement by visual contents," presented at The 7th International Conference on Web-Age Information Management (WAIM), Hong Kong, China, 2006.

[10] R. Zhao and W. Grosky, "Narrowing the Semantic Gap - Improved Text-Based Web Document Retrieval Using Visual Features," *IEEE TRANSACTIONS ON MULTIMEDIA*, vol. 4, pp. 189-200, 2002.

[11] Edward Chang, Kingshy Goh, Gerard Sychay, and G. Wu, "CBSA: Content-based soft annotation for multimodal image retrieval using Bayes Point Machines," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, pp. 26-38, 2003.

[12] J. V. R. J. L. Martinez-Fernandez, Ana M. Garcia-Serrano, Jose Carlos Gonzalez-Cristobal, "Combining textual and visual features for image retrieval," in *Accessing Multilingual Information Repositories*, vol. 4022: Springer Berlin / Heidelberg, 2006, pp. 680-691.

[13] X. H. Deng Cai, Zhiwei Li, Wei-Ying Ma and Ji-Rong Wen, "Hierarchical clustering of WWW Image search results using visual, textual and link Information," presented at the 12th annual ACM international conference on Multimedia, New York, NY, USA, 2004.

[14] P. H. Romaric Besancon, Pierre-Alain Moellic, Christian Fluhr, "Cross-media feedback strategies: merging text and image information to improve image retrieval," in *Multilingual Information Access for Text, Speech and Images*, vol. 3491: Springer Berlin / Heidelberg, 2005, pp. 709-717.

[15] D. Z. Tjondronegoro, J; Gu, J; Nguyen, A.; Geva, S, "Integrating Text Retrieval and Image Retrieval in XML Document Searching.," presented at INEX 2005., 2006.

[16] A. Tapus, S. Vasudevan, and R. Siegwart, "Towards a Multilevel Cognitive Probabilistic Representation of Space," *Proceedings of SPIE*, 2005.

[17] W.-Y. Ma and H. J. Zhang, "Benchmarking of image features for content-based retrieval," *Signals, Systems & Computers, 1998. Conference Record of the Thirty-Second Asilomar Conference*, vol. 1, pp. 253 - 257, 1998.

[18] D. Chai and A. bouzerdoum, "A Bayesian approach to skin color classification in YCbCr colorspace," presented at TENCON 2000, Kuala Lumpur, Malaysia.

[19] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image processing using MATLAB*. Upper Saddle River, N. J. :: Pearson Prentice Hall,, c2004.

[20] A. H. Kam, T. T. Ng, N. G. Kingsbury, and W. J. Fitzgerald, "Content based image retrieval through object extraction andquerying," *Content-based Access of Image and Video Libraries, 2000. Proceedings. IEEE Workshop*, pp. 91-95, 2000.

[21] R. O. Duda and P. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Commun. ACM*, vol. 15, pp. 11-15, 1972.

[22] J. C. v. Gemert, J.-M. Geusebroek, C. J. Veenman, C. G. M. Snoek, and A. W. M. Smeulders, "Robust scene categorization by learning image statistics in context," presented at CVPR Workshop on Semantic Learning Applications in Multimedia, New York, USA, 2006.

[23] D. Zhang and G. Lu, "Evaluation of similarity measurement for image retrieval," presented at Neural Networks and Signal Processing, 2003.

[24] Google Image Labeler, http://en.wikipedia.org/wiki/Google_Image_Labeler, Accessed: March 17, 2007.