A Unified Approach to Indexing Multimedia on the Web

Lakshman Jayaratne Zhuhan Jiang Athula Ginige
School of Computing and Information Technology
University of Western Sydney
Sydney, Australia
+61 2 46203143 +61 2 96859336 +61 2 46203639
k.jayaratne@uws.edu.au zhuhan@cit.uws.edu.au a.ginige@uws.edu.au

ABSTRACT
Indexing multimedia Web documents can be regarded as an important part of Web engineering, a concept first proposed [19] by one of the authors and his collaborators in 1998 at the World Wide Web WWW7 conference in Brisbane, Australia. Content-based indexing of multimedia has always been a challenging task. The enormity and diversity of the multimedia content on the World Wide Web (WWW) adds another dimension to this challenge. Today multimedia elements are increasingly being embedded in Web documents, and are being actively used to enhance the description of the document content. Since such documents over the WWW provide a rich source of information, the use of multimedia elements in Web documents has become very prevalent. In this paper, we first give a thorough review on the existing literature related to the traditional content-based image retrieval (CBIR) systems along with the methodology of relevance feedback. We then propose a unified approach for image indexing and retrieval for our *Image Search* retrieval system that performs relevance feedback on both the images' semantic contents represented by parts of the Web document as well as the low-level visual features. In addition, we will establish an approach with which semantic content and low-level features can be seamlessly integrated for the relevance feedbacks. More specifically, we will examine closely a number of ways that would combine visual and textual information for the content based indexing of multimedia on the Web. In particular, we will also propose and scrutinize different strategies of incorporating various mono media indexing approaches to create a multimedia indexing scheme for the purpose of image searches.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering, query formulation, relevance feedback, retrieval models, search process;* H.2.4 [**Database Management**]: Systems – *multimedia databases*

Keywords
Relevance feedback, image semantics, image retrieval, multimedia database.

1. INTRODUCTION
Digital images have become increasingly indispensable in the world we live today, with the Internet offering a very rich variety of such both interesting and valuable images. However, managing such images to facilitate their retrieval is still under intensive research and is far from being perfect. To search for relevant images from such a large collection of images over the Web calls for a decisive mechanism that would exploit fully the semantics of the images. Unfortunately, most image search engines on the Web fail to exploit these image semantics, resulting poor recall and precision performance. It is thus one of our purposes here to propose an image representation model on a unified approach that would perform the relevance feedback on both the images' semantic content and the low level visual features such as color, shape and texture.

Traditional CBIR systems [10, 6, 5] capture the visual content of an image such as color, shape and texture as its semantics and use these features as the basis for similarity matching. Although content-based image retrieval is very desirable in many applications, a full-fledged design and implement is extremely difficult. The ease with which humans capture the image content has so far not been fully understood, nor successfully automated. The main difficulties lie in extracting features from the images that capture the perceptual and semantic meanings, segmenting the images into regions corresponding to individual objects, and matching the images in a database with query image on the basis the extracted features. In addition, even though image retrieval systems allow a user to post queries for results, the retrieval accuracy is severely limited due to inherent complexity of describing chosen images exactly. On the other hand, visual content cannot effectively capture many useful image semantics, such like object, event, and relationship, and they do not scale well. Therefore, retrieval by visual content is still a long way from being matured.

With the increasing availability of digital images, automatic image retrieval tools are quested to provide efficient means to navigate through them. The relevance feedback approach to image retrieval [12, 16] in its original form is already a powerful technique, and has undergone active research over the past few years. Various adhoc parameter estimation techniques have subsequently been proposed for the relevance feedback. Relevance feedback offers a very important mechanism to improve the search accuracy. Typically, a system refines the query via feedback information from users to better and faster improve the subsequent retrievals. More recently, the need for a user to provide accurate initial queries has been reduced by improving the user's ideal query through the use of positive and negative example images dynamically selected by the user.

The drawback of these methods, however, is that they only perform relevance feedback on the low-level image features, and are thus unable to address the images' semantic content because the feedback is based entirely on the low-level image features such as color, shape and texture. If a user is searching for a specific object or event that cannot be sufficiently represented by available feature vectors, such relevance feedback systems will still not return many proper results even with a large number of user feedbacks. Consequently these systems work well only when the feature vectors can sufficiently capture the essence of the query.

In contrast, text-based systems [15, 2] use only keywords or free text description of images supplied by the users as the basis for the retrieval. These systems can be adopted for Web images since the textual content of the Web page in which the image is embedded provides the free text description. This is based on the observation that an image in a Web page is typically semantically related to its surrounding texts. These surrounding texts are used to illustrate some particular semantics of the image content, i.e. what objects are in the image, what is happening and where the place is. In particular, in a Web document, certain components are expected to provide more semantic information than other portion of the text. Therefore, in order to be able to search for relevant images among such a large collection of Web images we have to find mechanisms that exploit the semantics of the images and take advantage of the semantic contents of the images in addition to the low-level visual features.

This paper is organized as follows. We will first in section 2 extensively review the existing literature on the image retrieval systems that are prominently related to our proposed methodology. Section 3 will then describe in great details our proposed image representation model, and then present the semantic based as well as the low-level based feature extractions, the refinement of the relevance feedback approaches and the further details of our work. Section 4 will describe our *Image Search* retrieval system that we have implemented up to now on the basis of the proposed method, and will provide some experimental evaluations to illustrate its effectiveness. The concluding remarks will finally be made in Section 5.

2. RELATED WORK AND LITERATURE

2.1 The First Generation CBIR Systems

The original content-based image retrieval systems, classified as the 'first generation', index images in terms of low-level features. Examples of such systems include the IBM Query By Image Content (QBIC) system [11], the Virage system [1], and the VisualSEEK system [8]. Each of these allows the user to specify a query in some ways based on the low-level features extracted by the system.

2.2 The Second Generation CBIR Systems

In recent years there has been a proliferation in CBIR systems [12, 16, 17], roughly in our category of 'second generation'. These systems deliberately hide the low level features from the user. Instead of specifying color, shape and texture combinations the user supplies an example image and query for the similar ones. The idea of introducing the mechanism of relevance feedback [12, 16] into image retrieval systems was first demonstrated in the MARS system [17] developed at the University of Illinois. The user will be given ranked results or an opportunity to identify positive and negative images if he/she is not yet satisfied with the current query results. Such selected images will then be used to further refine the query.

2.3 The Third Generation CBIR Systems

We believe that the next (third) generation of image retrieval systems will have to address the limitations of the second generation systems by combining the low-level image feature space with the higher-level semantic space. Query formulation can then be performed using these higher-level semantics, most probably by entering a list of keywords representing the semantic contents of the desired images, as well as low-level image features.

2.3.1 The Methodologies

The systems we mentioned at the previous subsection perform relevance feedback all via the low-level feature vectors and have been unable to make use of actual semantics for the images themselves. The inherent problem with these approaches is that the low-level features are often not as powerful, at least at the grand scale, in representing complete semantic content of images as the keywords in representing text documents. In other words, applying the relevance feedback approaches used in low-level feature based retrievals may not be successful as textual retrievals. The low-level features are thus more suitable to be employed at query's later refining stages. This is why we will propose later on incorporating semantics of the images using keywords in relevance feedback for image retrieval in addition to the use of low-level features. In this approach, semantic relevance between images is learnt from user's feedback and used to improve the retrieval performance. Also, our proposed method will integrate both semantics and low-level features into an effective relevance feedback process in a new and unified manner. Loosely structured, but otherwise vast collections of images from the Internet bring another dimension to this already challenging task. Hence one of our purposes in this work is to investigate such image indexing mechanisms on the WWW on the extensive and propose a unified approach with combined evidence of multiple perspectives.

2.3.2 Traditional Search Engines for Web Images

A number of Web image search engines have been built in recent years including both research prototypes and commercial ones. Among the former category are WebSeer [13], WebSEEK [9], WebHunter [14] and iFind [18]. Commercial Web text search engines such as AltaVista, Yahoo and Google also offer image search facilities. In what follows we will overview the main systems that have been studied in this respect and the features they employed to index images. We will also show how they resemble or differ from our proposed work.

2.3.3 Recent Search Engines for Web Images

The iFind image retrieval system developed at Microsoft Research China [18] implements a relevance feedback approach via both higher-level semantics and certain low-level visual features. Since the textual semantic part is the closest to our later proposed system, we will explain below in some details its strategies and performance, as well as other possible significant aspects that are not yet developed. In this sense, our new system will be designed to develop and incorporate these new productive aspects, along with the additional lower level visual features.

"Giving Meanings to WWW images" [7] is another different approach to identify the semantics of an image within a Web document. The authors there presented a new model to represent

the content of images embedded in Web pages. In comparison with our proposed new model, this system however performs only relevance feedback on the high-level semantics. It is based on the observation that certain textual portions are expected to provide more semantic information of the image content in terms of, for example, what objects are in the image, what is happening and where.

Looking in other perspectives, there is a web-based image search agent called Diogenes [20], which is involved with the content-based indexing of person images on the Web. It retrieves Web pages and associates a person name with each facial image on those pages. This system also bears resemblance to our proposed method, in that it uses both higher-level semantic features and low-level visual features to obtain relevant images from the Internet, except that its prime design is to identify faces by using a training set of person images to find who appears in each image. It has a face detection module that examines the images on the Web page for human faces, a face recognition module that identifies the face by using a database of known person images, and a text/HTML analysis module that analyzes the body of the text for the clues about who appears in each image. The Diogenes search agent is one of the examples that have taken advantages of both textual and visual clues.

3. THE PROPOSED METHOD

There are two key issues that must be addressed in the design of our new retrieval system for Web images. First we need to determine a representation model for Web images based on the keywords and visual content on the image embedded in Web documents, as well as on the query semantics. Second we need to establish a similarity measure between an image and a query based on their representations. In this section, we shall address the first issue in great details.

3.1 Semantics of an Embedded Image

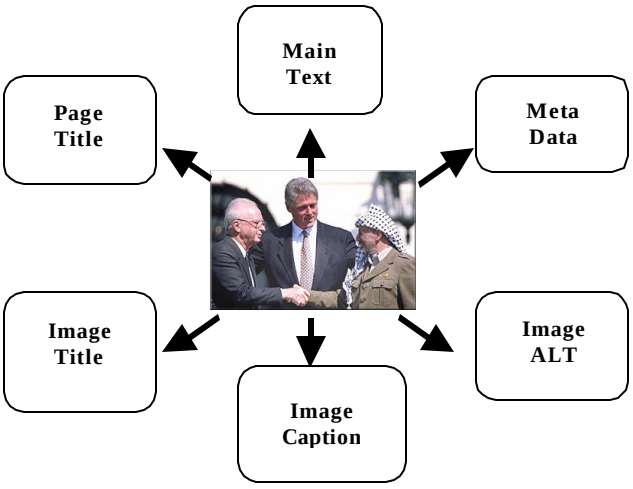

Figure 1: Image Semantic Representation

To understand the relationship between an image embedded in a Web document and its surrounding text, we conducted a preliminary study on a collection of images obtained from Web documents. Based on the relationship between an image embedded in Web document and its keywords, see Figure 1, we identified some parts of the textual content that are well related to the embedded image. These are

- *Image title* - Image title is basically a single string that indicates the main object the image is concerned with. For this purpose, a substring of image title or the image URL is assigned a higher significance.
- *Image alternate text* - The alternate text identified by the "ALT" property tag generally serves as a suitable textual replacement for an image or a descriptive phrase that typically represents an abstract of the image semantics.
- *Image caption* – The image caption often provides most of the semantics about an image. The caption includes the image's surrounding text in the corresponding Web document. It ranges from one sentence to a complete paragraph. For instance, a caption for an image is frequently kept in the same HTML table on the same column of the adjacent rows.
- *Page title* - Since images are often used to enhance a Web page's content, the page title is most probably related to the image's semantics. It usually gives a short sentence that best summarizes content of the Web page.
- *Main text* - This is based on the observation that an image in a Web page is typically more semantically related to its keywords.
- *Meta data* – HTML meta data may also provide additional information about the images in the document, such as the document description and headline stories.

3.2 Semantic Based Feature Extraction

The design of our text engine will employ a few direct techniques to associate keywords with the images and to retrieve images through the use of the keywords. For this purpose, we shall parse the Web document and collect the keywords for the corresponding images. More precisely, the text will pass through different processing stages. The first stage removes all the words that are so common in the language that they no longer provide any particular information about the content of the images. The second stage (stemming) extracts the root of the keywords from the candidate keywords. In this phase the system will typically remove the suffixes like the "s" of the plural form or the "ed" of the past tense.

Since the weight associated to each keyword represents the degree of relevance in which this keyword describes the image's semantic content, we have to implement a method of voting scheme that will determine the weights of the keywords within the Web document, those in the title, Meta tags, the caption of the image, the title of the image, image ALT string and keywords in the surrounding text of the image. By allocating higher weights for the keywords, especially those in the caption, title, and ALT string of the image we can discriminate the images from one another, in particular for the images from single Web document. From this image database with keywords we can then create an inverted file that lists the keywords against images with the degree of relevance, see Figure 2 below.

	Img_1	Img_2	Img_3	.....
Kw_1	W_{11}	W_{12}	W_{13}	.....
Kw_2	W_{21}	W_{22}	W_{23}	.....
Kw_3	W_{31}	W_{32}	W_{33}	.....
.	.			
.	.			

Figure 2: Inverted File - Keywords against images with the degree of relevance (Weights)

3.3 Semantic Based Relevance Feedback

Semantic based relevance feedback can be performed relatively easily compared to its low-level feature counterpart. It is basically a simple voting scheme, which updates the weights of the keywords against images without any user intervention. The process will in fact take the following steps.

1. Collect the user query keywords.
2. Compute the similarity of the images to the entered keywords to obtain the query results.
3. Identify positive and negative images from the query results.
4. For each positive image, check to see if any query keyword is linked to it. If so, increase the weight by a certain pertinent amount.
5. For each negative image, check to see if any query keyword is linked to it. If so, decrease the weight by a certain pertinent amount.
6. Show new results and go to step 3.

It can be easily seen that through this voting process, the keywords that represent the actual semantic content of each image will gradually receive a larger weight after user interactions. In this context of query refinement, relevance feedback introduces a learning mechanism to extract the discriminating features for Web images that would help the system to discriminate the images from one another for a given user query in order to achieve more accurate results.

3.4 Integration with Relevance Feedback based on Low-Level Features

Once images are available, we may also capture the visual content of the images, such as colors, shapes and textures, as part of the semantics, and use these features as the basis for similarity matching. This is to combine mono media indexing features to create a multimedia indexing scheme for relevance feedback.

The integration of textual semantics and low-level visual features can thus be carried out in different ways. In this connection, there will be a number of technical methods and strategies involved with establishing such as the similarity measurement. Their detailed expositions will out stretch both the scope and the allowed length of this paper. We will thus address these in a further publication.

4. EXPERIMENTAL EVALUATIONS

We have presented a unified approach in which high-level semantic and low-level feature based feedback can work together to achieve greater retrieval accuracy. In this section, we will describe the web-based image retrieval system *Image Search* that we have implemented up to now using this approach and illustrate some experimental evaluations and its effectiveness.

The *Image Search* retrieval system implements the unified approach presented in this paper. It is a Web based retrieval system in which multiple users can perform retrieval tasks simultaneously at any given time.

The *Image Search* system supports two modes of interaction: keyword based search, as well as search by example images. When a user enters a keyword-based query, the system invokes the query to search the images discussed in Section 3.3. The result page is shown in Figure 3.

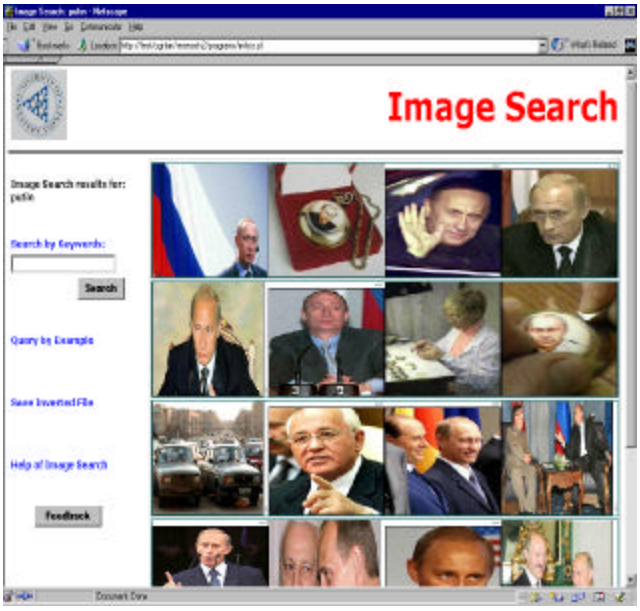

Figure 3: A snapshot of Image Search results for Vladmir Putin

The user is able to select multiple images from this page and click on the "Feedback" button to give positive and negative feedback to our system. New search results will then be presented to the user as soon as the "Feedback" button is pressed. At any point during the retrieval process, the user can select an image and click

on the "Query by Example" button to perform an example based query. One point of detail to note is that if the user enters a set of query keywords but some of these cannot be found in the inverted file, the system will simply output the images in the database corresponding to the keywords in the inverted file to let the user browse through and select the relevant images to feedback into the system.

The system presents 20 images for each query. The images are retrieved using the algorithm outlined in section 3. We are interested in examining how inverted file evolves with an increasing number of user feedbacks, we select a very clean, but from different category, image set as our starting point for the keyword-based search. The dataset that we have chosen is from the BBC and CNN Web sites. The current system is still being further extended and implemented whose full coverage is beyond the scope and length of this current paper. We will thus follow this up in great details in the near future.

5. CONCLUSION

In this work, we have presented a through review on the existing literature related to the traditional content-based image retrieval systems, and to the methodology of relevance feedback, and then proposed a approach that performs relevance feedback on both the images' semantic contents represented by the textual content of some parts of the Web document that are well related to the image's semantic content and the low-level features such as color, shape, and texture. The key in our new system is the integration of the semantics of the image with multiple keywords and the visual features (feature vector) with relevance feedback approach. We argue that combining these two semantics and allow them to benefit from each other yields a great deal of advantage in terms of both the retrieval accuracy and ease to use the system.

In contrast to the existing literature in this field, we have developed a method to construct an inverted file and to use a simple similarity matching technique to learn from the user queries and feedbacks to further improve this inverted file. Moreover, we have also proposed a new approach in which higher-level semantics and low-level feature based relevance feedbacks are combined to help each other in achieving higher retrieval accuracy with lesser number of feedback iterations required from the user.

6. REFERENCES

[1] A. Gupta and R. Jain, Visual Information Retrieval. Communications of the ACM, 40(5):71-79, May 1997.

[2] A.E. Cawkell, Imaging systems and picture collection management: a review, Information Service & Use, Pages 301-325, 1992.

[3] Anil K. Jain, Aditya Vailaya, Shape-based Retrieval: A case study with trademark image databases, Department of Computer Science, Michigan State University, East Lansing, Michigan.

[4] Anil K. Jain, Aditya Vailaya. Image Retrieval using Color and Shape. Department of Computer Science, Michigan State University, East Lansing, Michigan.

[5] Dobie M, Tansley R, Joyce D, Weal M. Lewis P, Hall W, MAVIS 2: A new approach to Content and Concept Based navigation, Multimedia Databases and Mpeg-7 (Ref. No 1999/056).

[6] Gudivada V.N., Jung G.S., An Algorithm for Content-Based Retrieval in Multimedia Databases, Multimedia Computing and Systems, 1996, Proceedings of the Third IEEE International Conference, 1996, Pages: 193-200.

[7] Heng Tao Shen, Beng Chin Ooi, Kian-Lee Tan, Giving Meanings to WWW Images, ACM Multimedia 2000, Los Angeles California, 2000.

[8] J. Smith, S. Chang, Intelligent Multimedia Information Retrieval, chapter Querying by colour regions using the VisualSEEK content-based visual query system, pages 23-41. AAAI Press, 1997.

[9] J.R. Smith, S.F. Chang, Visually Dearching the Web for Content, IEEE Multimedia, 4(3): 12-30, July-September 1997.

[10] Jian-Kang Wu, Content-Based Indexing of Multimedia Databases, Knowledge and Data Engineering, IEEE Transactions on, Volume: 9 Issue: 6, Pages: 978-989.

[11] M. Flickner, H. Sawnhey, W. Niblack, J. Ashley, Q. Huand, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Retkovic, D. Steele, P. Yanker, Intelligent Multimedia Information Retrieval, chapter Query by image and video content: The QBIC system, Pages 8-22, AAAI Press, 1997.

[12] Marinette BOUET, Chabane DJERABA, Visual Content-based Retrieval in an Image Database with Relevance Feedback, 1998 IEEE International Conference.

[13] Michael J. Swain, Charles Frankel, Vassilis Athitsos, WebSeer: An Image Search Engine for the World Wide Web, Technical Report TR-96-14, University of Chicago, Department of Computer Science, July 1996.

[14] Olaf Munkelt, Oliver Kaufmann, Wolfgang Eckstein, Content-Based Image Retrieval in the World Wide Web: A Web Agent for Fetching Portraits, In Proceedings of SPIE Vol. 3022, Pages 408-416, 1997.

[15] S. Al-Hawamdeh, B.C. Ooi, R. Price, T.H. Tng, Y.H. Ang, L. Hi, Nearest Neighbor Searching in a Picture Archival System, In Proceedings of ACM International Conference on Multimedia and Information System, Pages 17-34, 1991.

[16] Xiang Sean Zhou, Thomas S. Huang, Image Retrieval: Feature Primitives, Feature Representation, and Relevance Feedback, 2000 IEEE International Conference.

[17] Y. Rui, T. Huang, S. Mehrotra, M. Ortega, A Relevance Feedback Arichitectire in Content-Based Multimedia Information Retrieval Systems. In Proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries, 1997.

[18] Ye Lu, Chunhui Hu, Xingquan Zhu, HongJiang Zhang, Qiang Yang, A Unified Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems, ACM Multimedia 2000, Los Angeles California, 2000.

[19] Yogesh Deshpande, Athula Ginige, Steve Hansen, San Murugesan, Consolidate Web Engineering as a Discipline, World Wide Web WWW7 Conference, Brisbane, Australia, 1998.

[20] Yuksel Alp Aslandogan, Clement T. Yu, Evaluating Strategies and Systems for Content-Based Indexing of Person Images on the Web, ACM Multimedia 2000, Los Angeles California, 2000.