HIGH-LEVEL SEMANTICS OF IMAGES IN WEB DOCUMENTS USING WEIGHTED TAGS AND STRENGTH MATRIX

P.Shanmugavadivu[1], P.Sumathy[2], A.Vadivel[3]

[12]Department of Computer Science and Applications, Gandhigram Rural Institute,
Dindigul, Tamil Nadu, India
[12]psvadivu@yahoo.com, sumathy_bdu@yahoo.co.in

[3]Department of Computer Applications, National Institute of Technology, Trichy India
vadi@nitt.edu

ABSTRACT

The multimedia information retrieval from World Wide Web is a challenging issue. Describing multimedia object in general, images in particular with low-level features increases the semantic gap. From WWW, information present in a HTML document as textual keywords can be extracted for capturing semantic information with the view to narrow the semantic gap. The high-level textual information of images can be extracted and associated with the textual keywords, which narrow down the search space and improve the precision of retrieval. In this paper, a strength matrix is being proposed, which is based on the frequency of occurrence of keywords and the textual information pertaining to image URLs. The strength of these textual keywords are estimated and used for associating these keywords with the images present in the documents. The high-level semantics of the image is described in the HTML documents in the form of image name, ALT tag, optional description, etc., is used for estimating the strength. In addition, word position and weighting mechanism is also used for further improving the association textual keywords with the image related text. The effectiveness of information retrieval of the proposed technique is found to be comparatively better than many of the recently proposed retrieval techniques. The experimental results of the proposed method endorse the fact that image retrieval using image information and textual keywords is better than those of the text based and the content-based approaches.

KEYWORDS

Multimedia Information Retrieval, Web Image Retrieval, High-level Features, Textual Keywords

1. INTRODUCTION

The revolutionized advent of internet and the ever-growing demand for information sprawled in the World Wide Web has escalated the need for cost-effective and high-speed information retrieval tools. Many attempts have been made to use image contents as a basis for indexing and images retrieval. In early 1990, researchers have built many image retrieval systems such as QBIC [1], Photobook [2], Virage [3], Netra [4] and SIMPLIcity [5] etc., which are considered to be different from the conventional image retrieval systems. These systems use image features such as color, texture, shape of objects and so on whereas the recently devised image retrieval systems use text as well as image features. In the text based approach, the images are manually annotated by text descriptors and indexed suitably to perform image retrieval. However, these types of systems have two major drawbacks in annotating the keywords. The first drawback is

DOI : 10.5121/ijcsea.2011.1512

that a considerable level of human intervention is required for manual annotation. The second one is the inaccuracy annotation due to the subjectivity of human perception.

To overcome the aforesaid drawbacks in text-based retrieval system, content based image retrieval (CBIR) has been introduced [6]. However, these systems are suitable only for domain-specific applications. The low-level features such as color, texture and shape are efficiently used for performing relevant image retrieval [7] - [8]. Color histograms such as Human Color Perception Histogram [9] - [10] as well as color-texture features like Integrated Color and Intensity Co-occurrence Matrix (ICICM) [11] - [12] show high precision of retrieval in such applications. However, the inevitable semantic gap that exists between low-level image features and the user semantics drift the performance of CBIR still far from user's expectations [13]. In addition, the representation and storage of voluminous low-level features of images may be a costly affair in terms of processing time and storage space. This situation can be effectively handled by using keywords along with the most relevant textual information of images to narrow down the search space.

For this purpose, initially it is essential to explore the techniques to extract appropriate textual information from associated HTML pages of images. Many research methods have been proposed on the content structure of HTML document, including image title, ALT tag, link structure, anchor text and some form of surrounding text [20]. The main problem of these approaches is that the precision of retrieval is lower. This disadvantage has triggered the task of developing adaptive content representation schemes, which can be applied to a wide range of image classification and retrieval tasks. Further, the design techniques are also required to combine the evidences extracted from text and visual contents with appropriate mechanisms to handle large number of images on the web. Many recent techniques classify the images into one or more number of categories by employing learning based approach to associate the visual contents extracted from images with the semantic concept provided by the associated text. The principal challenge is to devise an effective technique to retrieve web-based images that combine semantic information from visual content and their associated HTML text. This paper proposes a faster image retrieval mechanism, which is tested on a large number of HTML documents. For this purpose, the HTML documents are fetched, using a web crawler. The content of the HTML documents is segregated into text and images and HTML tags. From the text, keywords are extracted and these keywords are considered to be the relevant keywords to represent the high level semantics of the images contained in the same HTML document.

In the following sections of the paper, the related works are presented. In section 3, the proposed method is elaborated along with a suitable example. The experimental result is presented in section 4. The conclusion is given in the last section of the paper.

2. RELATED WORKS

The retrieval of images from the Web has received a lot of attention recently. Most of the early systems have employed text based approach, which exploits how images are structured in web documents. Sanderson and Dunlop [16] were among the first to model image contents using a combination of texts from associated HTML pages. The content is modelled as a bag of words without any structure and this approach is found to be ineffective for indexing. Shen *et al* [14] have built a chain related terms and used more information from the Web documents. The proposed scheme unifies the keywords with the low-level visual features. The assumption made in this method is that some of the images in the database have been already annotated in terms of short phrases or keywords. These annotations are assigned either using surrounding texts of the images in HTML pages or by speech recognition or manual annotations. During retrieval, user's feedback is obtained for semantically grouping keywords with images. Color moments, color histograms, Tamura's features, Co-occurrence matrix features and wavelet moments are extracted for representing low-level features. Keywords in the document title of HTML page and image features have been combined for improving the retrieved documents of news category [17]. In this technique, from a collection of 20 documents chosen from one of the news site has been used and 43 keywords along with HSV based color histogram are constructed. While constructing histogram, saturation and hue axes are quantized into 10 levels to obtain H×S histogram with 100 bins. However, this technique is found to perform well for a small number of web pages and images only. In general, image search results returned by an image search engine contain multiple topics and organizing the results into different semantic clusters may help users. Another method has been proposed for analyzing the retrieval results from a web search engine [20]. This method has used Vision-based Page Segmentation (VIPS) to extract semantic structure of a web page based on visual presentation [18]. The semantic structure is represented as a tree with nodes, where every node represents the degree of coherence to estimate the visual perception.

Recently, a bootstrapping approach has been proposed by Huamin Feng *et al* (2004), to automatically annotate a large number of images from the Web [20]. It is demonstrated that the co-training approach, combines the information extracted from image contents and associated HTML text. Microsoft Research Asia [21] is developing a promising system for Web Image Retrieval. The purpose is to cluster the search results of conventional Web, so that users can find the desired images quickly. Initially, an intelligent vision based segmentation algorithm is designed to segment a web page into blocks. From the block containing image, the textual and link information images are extracted. Later, image graph is constructed by using block-level link analysis techniques. For each image, three types of representations are obtained such as visual feature based representation, textual feature based representation and graph based representation. For each category, several images are selected as non representative images, so that the user can quickly understand the main topics of the search results. However, due to index based retrieval, the time for processing is found to be on the higher side. Rough set based model has proposed for decompositions in information retrieval [22]. The model consists of three different knowledge granules in incomplete information system. However, while WWW documents are presented along with images as input, the semantic of images are exactly captured and thus retrieval performance is found to be lower. Hence, it is important to narrow down the semantic gap between the images and keywords present in WWW.

The textual information of WWW documents, which is the high-level semantics, can be effectively used for defining the semantics of the images without really using the low-level features. This kind of approach simplifies the semantic representation for fast retrieval of relevant images from huge voluminous data. This paper proposes a scheme for extracting semantic information of images present in WWW documents using only the textual information. The relationship between the text and images present in WWW documents estimated with frequency of occurrence of keywords and other important textual information present in image link URLs. Based on the experimental results, it is observed that the performance of the system is better than that of Google (www.google.com).

3. RELATED WORKS

3.1. Binary Strength Matrix using Keywords and Images

Let *H* be the number of HTML pages, *I* be the number of images and *K* be the set of keywords. Thus,

$H = \{h_1, h_2, h_3 \cdots h_n\}$, $I = \{i_1, i_2, i_3 \cdots i_m\}$ and $K = \{k_1, k_2, k_3 \cdots k_l\}$

where *n*, *m* and *l* denotes the total number of HTML pages, images and keywords respectively.

In order to effectively capture the semantics of *I* present in *H, K* can be used. The association between *K* and *H* can be written as:

$$stg(K \leftrightarrow H) = \left(\frac{stg(K)}{Max(stg(K))} \right) \qquad (1)$$

The above equation is the association between keywords to HTML pages and can be estimated using the frequency of occurrence of *K* in *H*. Since, *K* is the total number of keywords in a single HTML page 'h_p' may contain only 'k_q'of keywords. Now, the relation between each keyword 'k_j' where $(j = 1,2,3,\ldots,q)$ with a single HTML document can be written as

$$stg(k_j \leftrightarrow h_p) = \left(\frac{stg(k_j)}{Max(stg(k_j))} \right)_{j=1,2,\ldots,q} \qquad (2)$$

The above equation denotes the association between each keyword K_j in a single HTML document 'h_p'. Similarly, the relationship between a page and image can be derived. From Eq. 2, we get the importance i.e. strength of each keyword 'k_j' in a document 'h_p'. The strength is the measure of the frequency of occurrence and a keyword with a maximum frequency of occurrence is assigned higher strength value, which is 1. Similarly, all the keywords present in a particular document is assigned a strength value by which the importance of that particular keyword is estimated. The example depicted in Table 1 elucidates the estimation of strength using frequency of occurrence of a keyword in a HTML document.

Let the number of keywords in a HTML document is 5 and maximum frequency of occurrence (*FOC*) is 30.

Table 1. Strength Matrix using Frequency of occurrence

HTML Page	FOC	Strength
K_1	10	0
K_2	3	1
K_3	30	1
K_4	8	0
K_5	5	1

From the above example, we can observe that not all keywords are equally important. It is sufficient to consider only a set of keywords k_{stg} such that the strength of these keywords is greater than a threshold t_{stg}. In our approach, we have fixed this threshold as 25% of the maximum strength value. Now, it is important to estimate the strength of the keywords with the images. We have used Image Title, ALT tag, link structure and anchor text as high-level textual feature and a feature vector is constructed as given below

$$stg\left(k_j \leftrightarrow I_{hp}\right)=stg\left(k_j \leftrightarrow h_p\right)+m\left(TAG,k_j\right)+m\left(IN,k_j\right)+m\left(LS,k_j\right)+m\left(AT,k_j\right) \tag{3}$$

in above equation j = 1, 2, q and *m* is a string matching function with either 0 or 1 as the output. The output value of each component of above equation is in the rage of [0-1]. This relation is effective in capturing the importance of a keyword in a document and that of images. Both the strength value as well as image related textual information is combined to narrow down the semantic gap of image. Sample strength matrix with all features is depicted in Table 2.

Table 2. Strength Matrix using both frequency of occurrence and image high-level features

Keyword	FOC	$stg\left(k_j \leftrightarrow h_p\right)$	$\left[m(TAG,k_j)\right]$	$m\,(IN, k_j)$	$m\,(LS, k_j)$	$m(AT, k_j)$
K_1	1	0.033	1	0	0	0
K_2	30	1	1	1	1	0
K_3	20	0.66	1	0	0	1
K_4	8	0.26	1	1	1	0
K_5	5	0.16	1	0	0	1

In the above table, k_i is a keyword extracted from HTML documents. While extracting the keywords, the stop words are removed and the stemming operation is carried out. The high-level semantic information of the images can be extracted from the above table. Say for example, frequency of k_2 is high and also it is matching with most of the image related textual string and thus k_2 has more association with the HTML page and captures the semantics of the images.

$m(TAG,k_j), m(IN,k_j), m(LS,k_j)$ and $m(AT,k_j)$ are the match functions to match the similarity between images TAG and keyword, image name and keyword, link structures and keyword and Anchor Text with keyword respectively.

3.2 Weights to the Keyword Position

It is noticed from the above section that the entries in strength matrix is a binary value. While k_i is equal to any of the image related textual string, value 1 is assigned otherwise it is 0. Also, for any k_i, appearing around the images and any k_j appearing far from the image location compared to k_i is also treated equally (for $k_i = k_j$). Now, it is essential that both k_i and k_j (for $k_i = k_j$) should be assigned different values based on its position in the HTML document. In this work, the entire HTML page is segmented as various parts based on <img src > TAGs. In each partition, there is a set of keywords and associated position, which are used for assigning weights. As the notation followed in this paper, let $K = \{k_1, k_2, k_3 \cdots k_l\}$ be the keywords and $KP = \{kp_1, kp_2, kp_3 \cdots kp_l\}$ be the keyword position in the segment of HTML document. While k_i of a particular segment matches with any of the textual information in the <img src> TAG, more weight is assigned. Similarly, based on the physical position of a keyword, the weight assigned. The probability of a keyword k_i matches with any of the TAG information can be written as

$$TW = \Pr(k_i \mid ITAG(n)) \tag{4}$$

where $ITAG(n)$ is either $m(TAG, k_j), m(IN, k_j)$, or $m(AT, k_j)$. The value of *TW* is depends on the *ITAG(n)*. In this paper, based on our experience and analysis, the order of weights for the TAGs are $m(IN, k_j), m(TAG, k_j)$, $m(AT, k_j)$ and $m(LS, k_j)$. Say for example, in case, $m(IN, k_i) = true$, more weight is assigned to the keyword and for the case, $m(LS, k_i) = true$, less weight is assigned. Thus, weights are assigned for each TAG such that Image Name is given higher and Link State, ALT TAG, TAG are assigned lesser weight. Similarly, the keywords in a segment and corresponding distance are calculated based on its physical position. The weight of a keyword is calculated as below

$$KW = (k_i, kp_i) \tag{5}$$

where *i* is the total number of keywords in a segment. The function *KW* calculates the weight of a keyword with reference to its physical position from the image. Here, the reference point or position of a keyword is its physical position in that segment. Each keyword is referenced through a reference pointer and the distance from reference position to the keyword is considered as its index value. Higher the index value, lower the weight for the keyword and vice a versa. The final weight of a keyword for capturing semantics of an image in a segment is given as below

$$FKW = KW + (\Pr k_i \mid ITAG(n)) \tag{6}$$

4. EXPERIMENTAL RESULTS

For the purpose of crawling, the HTML documents along with the images, an internet crawler was also developed. The various stages of experimental setup is shown in Fig. 1

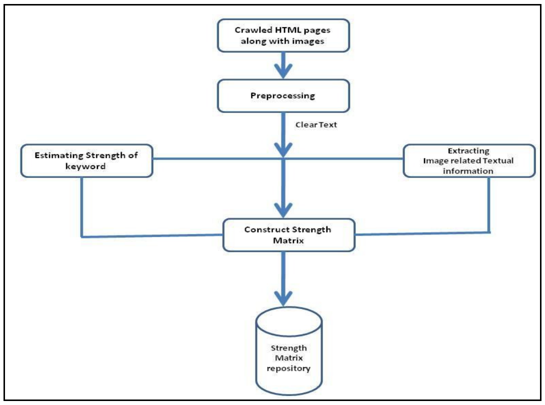
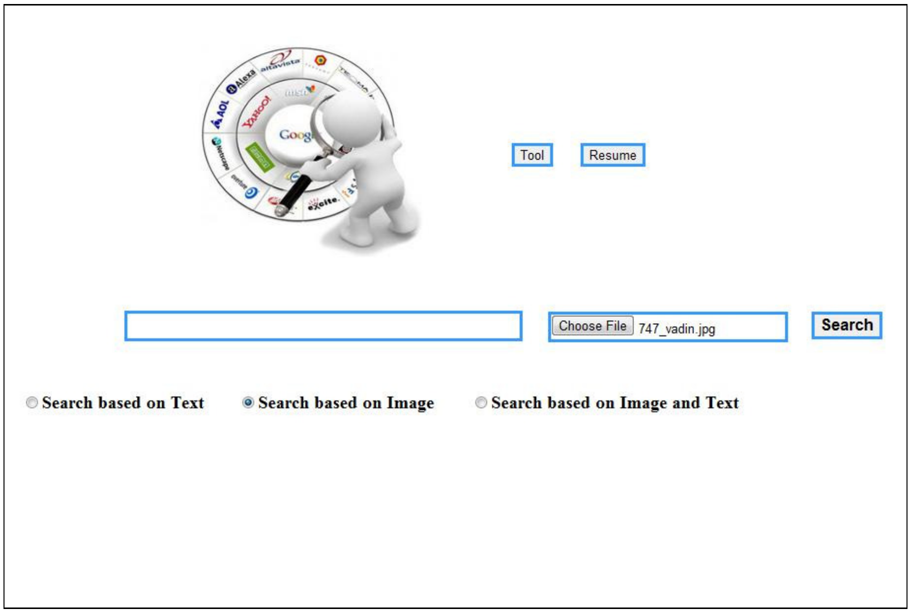

Crawled HTML pages along with images
Preprocessing
Clear Text
Estimating Strength of keyword
Extracting Image related Textual information
Construct Strength Matrix
Strength Matrix repository

Fig.1. Stages of Experimental Setup

The performance of the proposed technique is measured based on the crawled HTML pages along with the images from WWW. The textual keywords are extracted and the frequency of occurrence of all the keywords is calculated. The text information from URL link of images is also extracted. In addition, the page is segmented into various number of overlapping part. This process of segmentation is carried out for each image present in a HTML page using <img src> TAG>. The strength matrix is constructed using this information stored in a repository. During querying, the query keyword is found in the repository and based on the strength value, the result is ranked. In the experiment, many web documents from various domains such as sports, news, cinema, etc have been crawled. Approximately, 10,000 HTML documents with 30,000 images have been extracted and pre-processed for retrieval. The web crawler provides HTML pages along with the image content. The text in each page is categorized into normal text and image related text. While extracting the keywords, the stop words are removed and the stemming operation is carried out. The weighted matrix is constructed and it is stored in the repository using the clear text. For each page, this matrix is constructed and stored in the repository along with a relation between the document and image present. While making a query based on textual keywords, search will be carried out in the repository and the final ranking of retrieval is based on the weighted value. The images with higher weights are ranked first and will tip the retrieval result.

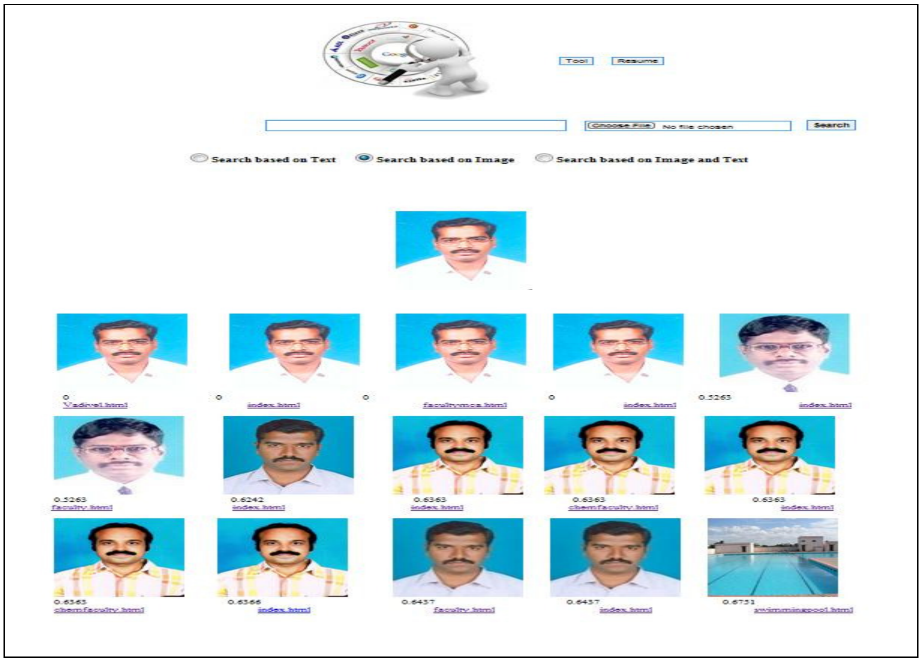


747_vadin.jpg

○ Search based on Text ● Search based on Image ○ Search based on Image and Text

Fig.2. The Query Interface of the Retrieval System Developed

In Fig.2, we present the query interface of the Multimedia Retrieval System Developed for measuring the performance of the proposed approach. The user interface can be used for using keyword and image as input. In this paper, we present the results only for the query in the form of text.



○ Search based on Text ● Search based on Image ○ Search based on Image and Text

0 Vadivel.html
0 index.html
0 facultymca.html
0 index.html
0.5263 index.html
0.5263 faculty.html
0.6242 index.html
0.6363 index.html
0.6363 chemfaculty.html
0.6363 index.html
0.6363 chemfaculty.html
0.6366 index.html
0.6437 faculty.html
0.6437 index.html
0.6751 swimmingpool.html

Fig.3. The Retrieval set of a given query from the Retrieval System

The output for a sample keyword for retrieving images is presented in Fig. 3. It can be observed that for a given query, relevant images are retrieved. Further, It is observed from Fig. 3 that the proposed system has retrieved the relevant image from WWW and ranked the relevant images higher. This is due to the fact that the strength matrix constructed from each page effectively captures the association between images and keywords. In addition, manually looked into the textual content of each HTML page and estimated the strength and are given for each image as the percentage of the estimated strength value was also computed manually. For evaluating the performance of the proposed approach in our system, the precision of retrieval is used as the measure. Moreover, the obtained results are compared with some of the recently proposed similar approach and are presented in Fig. 4. The average precision (P) in percentage for 10, 20, 50 and 100 nearest neighbors is given. We have compared performance with Hierarchical clustering (HC) [21]; Rough Set (RS) based approach [22] and Bootstrap Framework (BF) [20]. From the results shown in Fig. 4, it is observed that the performance of the proposed approach is quite encouraging. The precision of retrieval using the strength matrix is found to be high compared to others. The reason for this performance enhancement is due to the effectiveness of strength matrix in capturing the high level semantics of the images.

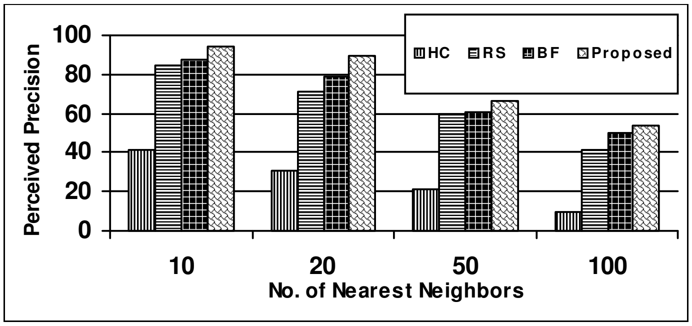


Fig 4. Comparison of Precision of Retrieval using strength matrix

It is well known that only the precision of retrieval alone is not sufficient for measuring the retrieval performance of any method. The Recall Vs. Precision is considered as one of the important measures for evaluating the retrieval performance. However, for measuring the recall value, it is important to have the ground truth. In this paper, we have measured the ground truth. For each HTML pages along with images, the distinct keywords present in that page are retrieved using a suitable SQL query. This gives us an idea about the distinct keywords present in a HTML page and used ground truth information. In addition, these distinct keywords are compared with the textual information in <img src> TAG for further acquiring ground truth information. Further, for all these keywords the physical position is also calculated for strengthening the ground truth. With the presence of the above mentioned ground truth, the recall and precision is calculated and the Recall Vs. Precision plot is shown in Fig. 5

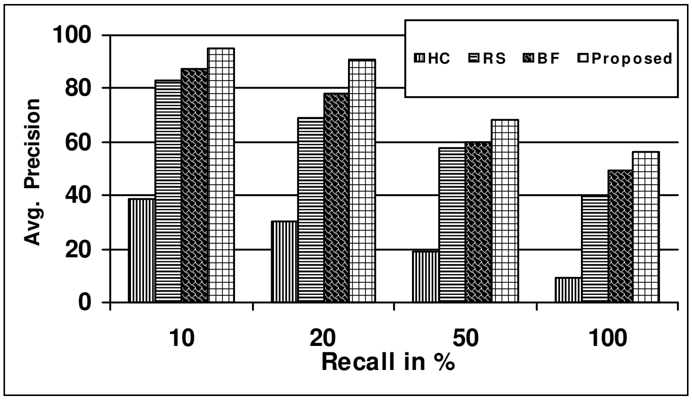


Fig 5. Comparison of Recall Vs. Precision of Retrieval

It can be observed from the above Figure is that the performance of the proposed approach is encouraging compared some of the similar recent approaches.

5. CONCLUSIONS

The role of textual keywords for capturing high-level semantics of an image in HTML document is studied. It is felt that the keywords present in HTML documents can be effectively used for describing the high-level semantics of the images present in the same document. Additionally, a web crawler was developed to fetch the HTML document along with the images from WWW. Keywords are extracted from the HTML documents after removing stop words and performing stemming operation. The strength of each keyword is estimated and associated with HTML documents for constructing strength matrix. In addition, textual information presents in image URL is also extracted and combined with the strength matrix. Based on the text category present in the <img src> TAG, weight is assigned. Similarly, the text position is also considered and weight is assigned. Finally, both of these weights are summed and final weight is calculated. It is observed from the experimental result that both textual keywords and keywords from image URL achieves high average precision of retrieval.

References

[1] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D.Petkovic & W. Equitz (1994) "Efficient and Effective Querying by Image Content", *Journal of Intelligent Information System*, Vol. 3, No.(3-4), pp. 231 – 202.

[2] A. Pentland, R.W. Picard & S. Scaroff (1996) "Photobook: Content-based manipulation for image databases", *International Journal of Computer Vision*, Vol. 18, No. , pp. 233–254.

[3] Gupta & R. Jain (1997) "Visual Information Retrieval", *Internal Journal of Communication of ACM,* Vol. 40, No. 5, pp. 70–79.

[4] W.Y. Ma & B. Manjunath, Netra(1997) "A toolbox for navigating large image databases" *In: Proceedings of International Conference on Image Processing*, pp. 568–571.

[5] J.Z. Wang, J. Li & G. Wiederhold (2001) "Simplicity: semantics-sensitive integrated matching for picture libraries", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No.9, PP. 947–963.

[6] Y. Liu, D.S. Zhang, G. Lu & W.-Y. Ma (2007) "A survey of content-based image retrieval with high-level semantics", *Pattern Recognition,* Vol. 40, No. 1, pp. 262–282.

[7] F. Long, H.J. Zhang & D.D. Feng (2003) "Fundamentals of content-based image retrieval" *Multimedia Information Retrieval and Management*, Springer, Berlin.

[8] Y. Rui, T.S. Huang & S.-F. Chang (1999) "Image retrieval. : Current techniques, promising directions, and open issues", *Journal of Visual Communication and Image Representation*, Vol. 10, No.4, pp. 39–62.

[9] Vadivel,A. Shamik Sural & Majumdar, A. K (2008) "Robust Histogram Generation from the HSV Color Space based on Visual Perception", *International Journal on Signals and Imaging Systems Engineering* ,Vol. 1, No.(3/4), pp.245-254.

[10] Gevers, T. & Stokman, H. M. G (2004) "Robust Histogram Construction from Color Invariants for Object Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol.26, No. (1), pp. 113-118.

[11] Vadivel, A. Shamik Sural & Majumdar, A. K. (2007) "An Integrated Color and Intensity Co-Occurrence Matrix", *Pattern Recognition Letters*, *Elsevier Science*, Vol. 28, No. (8), pp. 974-983.

[12] Palm, C. (2004) Color Texture Classification by Integrative Co-Occurrence Matrices. *Pattern Recognition*, Vol. 37, No. (5), pp. 965-976.

[13] Y. Liu, D. Zhang & G. Lu (2008) Region-based image retrieval with high-level semantics using decision tree learning, *Pattern Recognition* Vol. 41, pp. 2554-2570.

[14] H.-T. Shen, B.-C. Ooi & K.-L. Tan (2000) "Giving meaning to WWW images". *ACM Multimedia, LA, USA*. pp. 39-47.

[15] K. Yanai (2003) "Generic image classification using visual knowledge on the web", *ACM Multimedia, Berkeley, USA*. pp. 167-176.

[16] H.M. Sanderson & M.D. Dunlop (1997) "Image retrieval by hypertext links". *ACM SIGIR*, pp. 296-303.

[17] Zhao,R. & Grosky, W. I (2002), "Narrowing the Semantic Gap—Improved Text-Based Web Document Retrieval using Visual Features", *IEEE Transactions on Multimedia*, Vol. 4, No. (2), pp. 189-200.

[18] Cai, D. He, X. Ma, W-Y. Wen, J-R. & Zhang, H (2004) "Organizing WWW Images based on the Analysis of Page Layout and Web Link Structure", *In Proc. of International Conference on Multimedia Expo*, pp. 113-116.

[19] Cai, D. Yu,S. Wen, L.R. & Ma, W.Y. (2003) "VIPs a vision based page segmentation algorithm" *Microsoft Technical Report, MSR-TR-2003-79.*

[20] H. Feng, R. Shi, & T.-S. Chua (2004) "A bootstrapping framework for annotating and retrieving WWW images" *In: Proceedings of the ACM International Conference on Multimedia.*

[21] D. Cai, X. He, Z. Li, W.-Y. Ma & J.-R. Wen (2004) "Hierarchical clustering of WWW image search results using visual, textual and link information", *In: Proceedings of the ACM International Conference on Multimedia.*

[22] Chen Wu & Xiaohua Hu(2010) "Applications of Rough set decompositions in Information Retrieval". *International Journals of Electrical and Electronics Engineering* Vol. 4, No. 4, (2010).

Authors

Dr. P Shanmugavadivu is currently working as Associate Professor. Her rese arch interest includes image and video processing and analysis, Multimedia Information Retrieval

Mrs. P. Sumathy is Research Scholar in the Department of Computer Science and Applications of Gandhigram Rural Institute Dindigul, India. She is currently working as Assistant Professor in the Department of Computer Science of Bharathidasan Univers ity Trichy India. Her research interest is Multimedia Information Retrieval

Dr. A Vadivel is currently working as Associate Professor National Institute of Technology Trichy. His Research interest includes Multimedia Information Re trieval, Image and video processing and Analysis.