Image Retrieval Using Non-Binary-Weighted Approach

Khor Siak Wang†, Fatimah Ahmad††

†Faculty of Information Communication and Technology, Universiti Tunku Abdul Rahman, Petaling Jaya, Malaysia
††Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Malaysia

Summary
Colour histogram has been a popular technique for colour indexing and image retrieval. However, it suffers from a major drawback, i.e. fails to take into consideration the spatial information of an image during the image matching process. In this paper, we present a new technique for retrieving images having similar chromatic content in a query image. This is accomplished by assigning non-binary weights to index terms in the query and stored images, a popular technique making use of the vector model from the literatures of information retrieval for image similarity match. The index terms are derived from the image content. The proposed technique will be benchmarked against the histogram technique. A standard dataset that currently contains 1338 colour images, known as Uncompressed Colour Image Database (UCID), is used for the benchmarking purpose. With the system developed using both Visual Basic and Java, it has shown an improved performance in term of retrieval accuracy, with an averaged precision value of about 70 % over the traditional histogram technique that only enjoyed approximately 20 % of precision value for retrieving images having similar chromatic content with a query image.

Key words:
Vector Model, Hue Pattern, Bucket, Boolean Model, Vector Model, Probabilistic Model

1. Introduction

Images now play a crucial role in diverse fields [1] such as medicine, journalism, advertising, design, education, entertainment etc. The twentieth century has unquestionably witnessed an unparalleled growth in the number of images and availability of image retrieval tools. This has undoubtedly given a raised interest among researchers how images could be efficiently and accurately retrieved from huge collections of stored images. Apparently, with the advancements in digital technologies and devices, more and more digital images are becoming available every day. Many techniques have been proposed either at the experimental stage or commercial environment, to ease the image retrieval process. Of all these proposed techniques, one of the most popular and common techniques is based on the use of histogram, which is a representation of the global distribution of colors in an image, derived by counting the number of pixels having similar chromatic content. However, the use of histogram for image retrieval does suffer from a major problem, i.e. it does not take into account the spatial information of an image.

In this paper, a novel approach for colour image retrieval based on the vector model from the literatures of information retrieval has been implemented. The proposed technique takes into account the spatial data of an image and thus is proven to the more accurate when benchmarked against the histogram-based technique for retrieving images having similar chromatic content of a given query image. With this properly devised colour space discretisation technique, collections of pixels having similar colour content are identified, grouped and sorted according to the proportional spread, in term of total pixels in a given image. Weights are assigned to each identified groups. These term weights are ultimately used to compute the degree of similarity between each stored image in the system and the query image. By sorting the retrieved images in decreasing order of this degree of similarity, the vector model takes into consideration images, which match the query terms only partially.

The rest of the paper is organised as follows. Related works for colour-based image retrieval are given in Section 2. Some cursory glance of the classic models is given in section 3. Section 4 gives an overview of the proposed technique. Section 5 provides the details of the experimental setup to run the system. Results are reported in Section 6 with some concluding remarks and future works given in Section 7.

2. Previous Works

A technique known as *Integrated Colour-Spatial Approach* is proposed by Hsu et al. [2] where a set of representative colours are used to aid in the image matching process. After the representative colours have been determined, the spatial information about these colours are obtained using maximum entropy discretization with event covering method. Selection of colours is performed by using two colour histograms. Once the selection of colours is

Manuscript received September 5, 2009
Manuscript revised September 20, 2009

complete, each stored image will then be partitioned into regions of rectangular shapes with each region being occupied by a single colour, derived from the colour discretization process explained above. The partitioning process is performed with the help of stacks, implemented in C language that runs under the Unix environment with a test data of 260 images. To test the similarity between two images, a similarity distance function has been developed, which is based on the degree of overlap between regions of the same colour.

Unlike the technique proposed by Hsu et al. [2], Smith and Chang's algorithm [3] also partitions the image into regions but the authors allow a region to contain multiple colours. The partitioning process is making use of a sequential labeling algorithm. They extend the technique histogram back-projection [4] to back-project sets of colours on to the image. Instead of blurring the back-projected images, the authors use morphological filtering to identify the colour regions.

Pass and Zabih [5] have proposed a histogram-refinement approach for image matching process. More formally, their technique is known as Colour Coherent Vectors (CCV) where pixels with similar colour content will be connected to form larger components. From these connected components, a given pixel will then be judged whether it belongs to a given component. For instance, when a yellow pixel belongs to an identified component, the pixel is said to have high coherence. Otherwise, it's of low coherence. Once a pixel has been determined for its coherence, the information of the pixel will then be stored in vector format, ready for the retrieval process.

From all the techniques discussed above, they are either based on histogram-refinement method or image partitioning method. The colour correlograms [6] approach is neither of these methods. The colour correlogram is concerned with how the spatial correlation of pairs of colours changes with distance. A colour histogram captures only the global colour distribution within image whereas the colour correlogram considers also its local colour distribution.

Stricker and Dimai [7] proposed a technique known as Spectral Covariance and Fuzzy Regions, where an image is divided into five fuzzy with central region being the most "important" region during the image matching process. They proposed a similarity function that enables the query image to be compared with rotated versions of the images in the database at rotation angles 0, 90, 180 and 270 degrees.

The Spatial Colour Histogram [8] approach makes use of three types of spatial colour histograms, i.e. annular, angular and hybrid colour histograms with the intention of capturing spatial information of pixels within an image. The key difference between the proposed technique with the ones highlighted above is that both the spatial information and colour information are equally important and the spatial and colour contributions to a final histogram can be balanced through tuning of some parameters.

The methodology Spatial-Chromatic Histogram [9] considers the statistical aspects of pixels, i.e. both mean and standard deviation having the same colour and their spatial arrangement within the image. The authors synthesise some values information about the location of pixels having the same colour and their arrangement within the image.

To accurately retrieve colour images, the authors [10] use signature bit-strings approach together with an appropriate similarity metric. Experiment run on a heterogeneous database of 20,000 images demonstrated that the proposed technique could retrieve relevant colour images more accurately as compared to the well-known histogram-based approach.

By analyzing the color distribution of an image, the authors [11] present a new method known as metric histogram for content-based image retrieval. By taking into consideration the adjacent bins of histograms, the authors manage to demonstrate a reduction in dimensionality of the feature vectors extracted from images, which also lead to speedier retrieval process.

To properly describe the color variation of pixels within an image, the authors [12] propose a color-complexity image feature. They also present color-spatial feature to state the pixel color distributions on different locations in an image. Since these two features are highly complementary, they integrate them to provide an image retrieval system.

The authors [13] introduce two new descriptors, color distribution entropy (CDE) and improved CDE (I-CDE), to describe the spatial information of colors within an image for the purpose of image retrieval. In comparison with the spatial chromatic histogram (SCH) [9] which also measures the global spatial relationship of colors, the experiment results show that CDE and I-CDE outperforms SCH for colour image retrieval.

3.0 Classic Models

Generally, they are three classic models in information retrieval, namely, Boolean, Vector, and Probabilistic models and these classis models consider that each

document is described by a set of representative keywords called index terms [14]. An index term is simple a word whose semantics helps in remembering the document's main themes. Therefore, index terms are used to index and summarize the document contents. Similarly, for an image, the index term represents the proportion of pixels having similar hue pattern within an image. All the computed index terms will be weighted and the derived formula from the vector model is then used to compute the similarity between each stored image against the query image.

The Boolean model considers only either the presence or absence of index terms in a document, which could be too limiting whereas the Probabilistic model relies heavily on the use of probability theory, which could be complicated. The vector model recognizes the fact that exact match of stored documents and query documents is quite unlikely to occur [14]. Similarly, it is very unlikely that, given a collection of stored images, we can always find images that have perfect match with a query image. Therefore, the use of similarity function in the vector model, which is for image similarity match, is appropriate. The definition of vector model [14] is given below: -

> For the vector model, the weight $w_{i,j}$ associated with a pair (k_i,d_j) is positive and non-binary. Further, the index terms in the query are also weighted. Let $w_{i,q}$ be the weight associated with the pair $[k_i,q]$, where $w_{i,q} \geq 0$. Then, the query vector q is defined as $q = (w_{1,q}, w_{2,q}, w_{3,q}, w_{4,q}, w_{5,q}, \ldots, w_{t,q})$ where t is the total number of index terms in the system. As before, the vector for a document d_j is represented by $d_j = (w_{1,j}, w_{2,j}, w_{3,j}, w_{4,j}, w_{5,j}, \ldots, w_{t,j})$.

The similarity function is then defined as: -

$$sim(d,q) = \frac{\sum_{i=1}^{t} (w_{i,j} * w_{i,q})}{\sum_{i=1}^{t} w_{i,j}^2 \quad \sum_{i=1}^{t} w_{i,q}^2}$$

Since $w_{i,j} \geq 0$ and $w_{i,q} \geq 0$, *sim (d,q)* varies from 0 to +1. Therefore, rather than assessing whether an image is relevant or not, the vector model ranks the images according to their degree of similarity to the query image. In other words, an image might still be retrieved even if it matches the query image only partially. As an example, we can always establish a threshold on *sim (d,q)* and retrieve the images with a degree of similarity above that threshold.

4.0 Proposed Technique

Generally, the entire model (figure 1.0) for image retrieval of the proposed technique is divided into two phases, as follows: -

(a) Phase 1 - *Image Preprocessing*
(b) Phase 2 - *Image Retrieval*

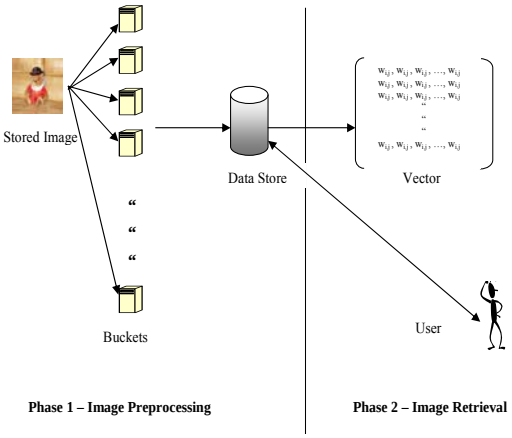

Figure 1.0: Two-Phase Image Retrieval

During phase 1, each stored image will be preprocessed where the chromatic content of the image will be analyzed and stored into a text file, which will then be fetched into memory store for image retrieval purpose. During this phrase, each and every pixel within a stored image will be grouped into a bucket depending on its hue pattern. In the research work, 24-bit colour images from the Uncompressed Colour Image Database (UCID) are used as testbed. With 24-bit colour depth value, the number of colours present in each image could be tremendously huge. Thus, for an efficient and effective computation of similarity match of images, a drastic reduction in the number of colours used to represent all the colours possibly present in an image is required [15]. To do that, the entire RGB colour space has to be discretized so that RGB for every pixel within an image can be completely represented. Each of the three primary colours is equally divided into three equally-spaced partitions and tagged as 1, 2 and 3. All the possible unique combinations of tags from these three primary colours are performed and the derived values are formally termed as hue pattern. With such discretization method, the maximum number of colours present in an image has now become 27, i.e. 111,

112, 113, 121, 122, 123, … ,333. Each of these hue patterns is tagged with a particular bucket, which will result in a total of 27 buckets used during the sorting process, in phase 1

In other words, all the pixels with the same defined hue pattern will be stored into a single bucket. The number of buckets depends on the number of defined hue patterns from the RGB colour space. Once all the buckets have been completely filled, i.e. when the entire image is completely processed, the collection of buckets will then be sorted in descending order of pixel quantities, i.e. ranging from the bucket that contains the most pixel count to the one with the least pixel count. Intuitively, dominant colours within an image can also be determined. During the sorting process, the pixel count within a bucket will be converted into a non-binary weight, i.e. the percentage value of its present rate over the image dimension. These weights will be stored into a text file.

During phase 2, a similarity function, taken from the classic information model [14], i.e. vector model will then be used for image similarity match. The computed non-binary weights (figure 1.1), which are stored in the text file during phase 1, serve as input data into the similarity function. Figure 1.1 shows an extract from the text file. It lists the first 20 scores of stored images having the closest chromatic match with a given query image. For instance, for the query image 00001, the first five stored images with the closest match are 00001, 00077, 00992, 01242 and 00626 with their respective non-binary weights as 1.0000, 0.9976, 0.9971, 01242 and 00626.

Figure 1.1: Computed non-binary Weights

The returned values from this function will then be used to judge the similarity of a retrieved image against the query image.

5.0 Experimentation

A standardized image database for benchmarking purpose is always the main weakness in the field of content based image retrieval (CBIR) [16][17] where authors involved in CBIR-related research always use their own preferred datasets for benchmarking purpose. This undoubtedly introduces biasness in their experiment runs. To solve this problem, authors from the Notthingham Trent University [16] have come out with a standardized set of colour image database, formally known as Uncompressed Colour Image Database (UCID).

At present, the dataset consists of 1338 uncompressed images on a variety of topics such as natural scenes and man-made objects, both indoors and outdoors.

To take the advantage of GUI-capability, Microsoft Visual Basic 6.0 is used to design the entire prototype system whereas the image matching process, which is more computationally intensive, is developed using Java. Also, twenty query images have been chosen for the benchmarking purpose and they form twenty queries. These twenty query images are also picked from the UCID dataset. These carefully selected images comprise colours in different compositions and locations within the images. These queries will be used to compare the retrieval performance, in term of retrieval accuracy using both the histogram technique and the proposed technique for colour-based image retrieval.

6.0 Discussions and Results

Generally, from the tabulated results, all the twenty queries run using the proposed technique have produced better retrieval accuracy. This is obvious when each of the *Recall and Precision* graphs plotted has generated higher precision values using the proposed technique.

The table below provides a summary of all the computed values of precisions for all the 11 standard recall values for images using both histogram approach and proposed technique respectively.

Standard Recall Value	Averaged Precision (20 Queries)	
	Histogram (%)	Proposed Technique (%)
0	32.1	100.0
10	30.2	95.9
20	28.6	93.9
30	27.0	88.7
40	27.0	86.1
50	26.8	75.2
60	24.4	70.6
70	17.4	69.4
80	8.7	40.2
90	6.3	32.1
100	5.1	30.2
Averaged Precision (11 Standard Recall)	21.2	70.6

Table 1.0: Averaged Precision Values for Twenty Queries

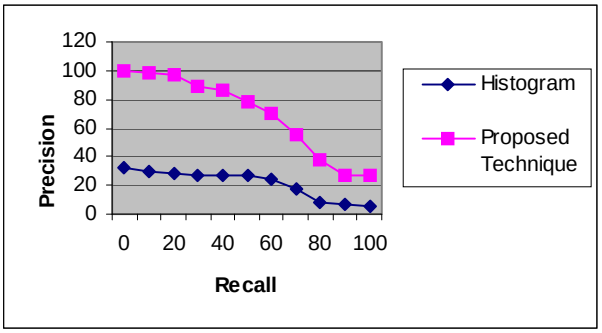

Figure 1.2: Averaged Recall & Precision Values

From the tabulated data from table 1.0, it is pretty clear that retrieving images using the histogram technique has produced considerably low averaged precision rates for each of the standard recall values. This is mainly due to the fact that the histogram approach does not take into account the spatial locations of colours present in an image. For instance, consider one of the query images used for benchmarking test, which is a flag with both red and white interleaved at the center of the image, has produced the query results as shown in both figure 1.3 and figure 1.4.

Figure 1.3: Results of the Query Image "Flag" Using Histogram Technique

Figure 1.4: Results of the Query Image "Flag" Using the Proposed Technique

In figure 1.2, the first few images are irrelevant (but ranked to be the most relevant to the query image by the histogram technique) but retrieved by the system. This is due to the fact that these images contain both the red and white with similar proportion as the query image but in scattered locations.

From this example, it is obvious that using histogram technique for colour image retrieval considers only the global colour of images where spatial locations of colours are entirely ignored.

7.0 Overall Conclusion and Future Works

The retrieval results from the proposed technique indicate that the proposed technique has outperformed the traditional histogram approach for retrieving images having similar chromatic content as a given query image. It is also important to note the significance of adopting the vector model, which was initially develop to be used for information retrieval, in retrieving images with good retrieval output. However, a number of limitations need to be considered. Firstly, depending on colour quantization technique used, the number of resulting hue patterns could be huge, which may impede the system performance. Secondly, the use of probability model has not been fully studied. It may yield more accurate results than the vector model.

Therefore, it would be interesting to assess the effect of the suitable number of hue patterns used in obtaining much better retrieval accuracy without affecting the system performance. Also, the applicability of the probability model may also be attempted in the near future.

References

[1] Ju Han, Kai-Kuang Ma. Aug 2002. "Fuzzy color histogram and its use in color image retrieval". IEEE Image Processing, IEEE Transactions on. Pages: 944-952.

[2] Hsu Wynne, T.S. Chua, and H.K. Pung. 1995 . "An integrated colour-spatial approach to content-based image retrieval." In ACM Multimedia Conference. Pages: 305-313.

[3] Smith John and Shih-Fu Chang. February 1996. "Tools and techniques for colour image retrieval." SPIE Proceedings , 2670.

[4] Swain M. J. and D.H. Ballard. 1991. "Colour Indexing". Int. Journal of Computer Vision Vol. 7, No. 1. Pages: 11-32.

[5] Greg Pass, Ramin Zabih, Hustin Miller. Dec. 1996. "Histogram Refinement for Content-based Image Retrieval". In Proc. 3rd IEEE Workshop Applications Computer Vision." Pages: 96-102.

[6] Huang, S. R. Kumar, M. Mitra, W. J. Zhu and R. Zabih. 1997. "Image Indexing Using Colour Correlograms". In Proc. IEEE CVPR. Pages: 762-768.

[7] Stricker M. and A. Dimai. 1997. "Spectral covariance and fuzzy regions for image indexing". Machine Vision Application Vol. 10. Pages: 66-73

[8] Aibing Rao, Rohini K. Srihari, Zhongfei Zhang. 1999. "Spatial Colour Histograms for Content-based Image Retrieval". Center of Excellence for Document Analysis and Recognition, State University of New York At Buffalo.

[9] Cinque L., G. Ciocca, S. Levialdi, A. Pellicano, R. Schettini. 2001. "Colour-based Image Retrieval Using Spatial-Chromatic Histograms". Image Vision Computing. Pages: 979-986.

[10] Mario A. Nascimento, Vishal Chitkara, "Information access and retrieval: Color-based image retrieval using binary signatures,". Proceedings of the 2002 ACM symposium on Applied computing SAC '02. March 2002.

[11] A. J. M. Traina, C. Traina, J. M. Bueno, F. J. T. Chino, P. Azevedo-Marques, "Efficient content-based image retrieval through metric histograms, ". World Wide Web, Volume 6 Issue 2, June 2003.

[12] Yung-Kuan Chan, Chih-Ya Chen, "Image Retrieval System based on color-complexity and color-spatial features, ". Journal of Systems and Software, Volume 71 Issue 1-2. April 2004.

[13] Junding Sun, Ximin Zhang, Jiangtao Cui, Lihua Zhou, "Image retrieval based on color distribution entropy, " Pattern Recognition Letters, Volume 27 Issue 10. July 2006.

[14] Baeza-Yates, R. & Ribeiro-Neto, B. 1999. Modern Information Retrieval. New York: Addison-Wesley. Pages 27-30

[15] Chiou-Ting Hsu Chuech-Yu Li. Oct, 2005. "Relevance Feedback Using Generalized Bayesian Framework With Region-based Optimization Learning". Image Processing, IEEE Transactions On. Volume 14, Issue 10.

[16] Gerald Schaefer and Michal Stich. 2003. "UCID – An Uncompressed Colour Image Database". Schoold of Computing and Mathematics, The Nottingham Trent University, Nottingham, United Kingdom.

[17] Jiang W. Er, G. Dai, Q. Gu, J. March 2006 "Similarity-based Online Feature Selection in Content-based Image Retrieval". Image Processing, IEEE Transactions On. Volume 15, Issue 3.