International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

Color Histogram Based Image Retrieval

Sivaraman.K, Asst. Professor Bharath University

Abstract

Color histograms are widely used for the Content-Based Image Retrieval (CBIR). In this paper we propose a new image retrieval technique that combines color and edge features for image indexing. We employ the YCrCb (luminance/red chrominance/blue chrominance) color space for edge histograms in our work. We use Euclidean distance for distance measurement between a query image and images in a database.

Keywords -color-based image retrieval, content based Image retrieval, Euclidean distance, histogram intersection.

Introduction

With the increasing popularity of image management tools such as Google's image search and photo album tools such as Google's Picasa project, as well as image search applications in general social networking environment, the quest for practical, effective image search in the web context becomes ever more important. The research community has seen a number of algorithms and tools that facilitate image retrieval. This paper examines one particular algorithm that is based on image retrieval. We implemented the algorithm in

Java and compare the effectiveness of the algorithm with other popular image search tools. We conclude that while the algorithm is effective, it needs to be fine tuned before being deployed as a practical tool. We offer some thoughts how this might be done.

Literature reviews

Currently the most popular search engines for images rely on the comparison of metadata or textual tags associated with the images. This methodology relies on human intervention to provide an interpretation of the image content so as to produce tags associated with the image. However, the ever increasing prevalence of large image databases has resulted in the development of algorithms to augment and replace tag based image retrieval with content based image retrieval. These algorithms compare the actual content of the images rather than text which has been annotated previously by a human being. There are a number of features that can be extracted from an image for comparisons based on their content. Indeed, the Photo book application developed at MIT allows users to perform image retrievals based on user developed models for various information extractions. Generally similarity between two images is based on a computation involving the Euclidean distance or histogram intersection between the respective extracted features of two images. Both these methods involve an intuitive extension of the mathematical definition of a distance between two objects. The three most common characteristics upon which images are compared in content based image retrieval algorithms are color, shape and texture.

Shape and texture based retrieval

Utilizing shape information for automated image comparisons requires algorithms that perform some form of edge detection or image segmentation. Segmentation refers to the identification of the major color regions in an image. These regions can then be compared from one image to the next. Edge detection tends to be slightly more complicated as it attempts to identify the major contours and edges in a given image. These edges may be then compared based on their direction, with respect to image edges. The advantages of this method include its applicability to black and white images. However, the performance of the algorithm is not invariant on scale or translation manipulations of images. Information regarding the texture of images can be even harder to extract automatically during retrieval. Generally algorithms rely on the comparison of adjacent pixels to determine the contrast or similarity between pixels

Color based retrieval

By far the most intuitive information that can be extracted from images for comparison is the color characteristics of an image. This paper attempts to explore and analyze such an algorithm that compares images based on their color content. A number of algorithms have been developed since the late 1980s that use color information extracted from images for retrievals [10]. A most basic form of color retrieval involves specifying color values that can be searched for in images from a database. Indeed, Google's image organization and editing software, Picasa 3.0, allows users to use an experimental tool to search for certain colors in images. Even this basic method presents challenges in implementation due to the different manners in which computers and human 'see' colors. Computers represent all visible colors with a combination of some set of base color components, generally Red, Green and Blue (RGB). Thus, images perceived by a computer to contain a large component of red may not necessarily appear 'reddish' as perceived by a human eye. Indeed, Picasa's experimental tool suffers from this and returns certain unintuitive results. Other image retrieval methodologies rely on specifying more precisely the nature of the color that is to be retrieved. The method offers numerous benefits with only a few limitations. Firstly, image retrieval based on this concept should accurately retrieve images despite the manipulation of orientation, size and position of a certain image. Also easy in terms of processing content information. A limitation of this algorithm is its inability to easily incorporate the spatial characteristics of the colors in an image. This is particularly true for images stored in Meta or Vector formats that contain more information than simply an array of pixels. Researchers from Stanford University have explored possible solutions by implementing vector quantization strategies that incorporate the distribution of colors in an image. Also, black and white images can necessarily not be compared using an algorithm based solely on color comparisons. In their paper, Jain and Vailaya further analyze this methodology. The image retrieval utilized during their experimentation computes similarity based on the similarity of three different histograms, one for each component of a RGB pixel. The similarity is computed using a Euclidean distance function comparing each 'bin' of the histograms. Retrieval was then carried out by searching for images with the minimum distance to a query image. The experiment carried out testing using an image database consisting of trademarks. The results demonstrate that even with this relatively simple implementation, over 90% of the time an image query is matched accurately with an image in the database. In addition, ignoring a pixel ranges not present in one of the images can reduce the impact of background color on the result. Their experiment was carried out on 500 images taken from the Simplicity content based image retrieval database using various implementations of the color histogram. The results from querying these databases with images were analyzed using precision versus recall graphs. Recall signifies the number of relevant images in the database that are retrieved in response to a query Precision refers to the proportion of the retrieved images that are relevant to the query. Thus, if precision can be increased without sacrificing recall the algorithm is performing well. The experiments analyses showed that the HSV model in conjunction with a histogram intersection method produced the most successful query responses.

Current content based image retrieval systems

Most existing platforms for retrieving images based on image content implement algorithms that extract a Combination of shape, texture and shape features from an image. Then weights are generally assigned to each piece of information extracted from the images and an overall similarity is computed. Images can then be ranked based on this similarity computation. A number of both closed and open source software products can be found. A popular system that has been implemented is IBM's QBIC system. The system has been implemented by the Hermitage Museum website which allows users to search through their digital library of artwork using QBIC's color and layout comparison tools. In addition, there are a number of other solely online application offering services that perform some form of content based image retrieval. Some of these applications were used in the testing process for color histogram technique. The final such system utilized during our testing is AIRS (Advanced Image Retrieval) developed by the Corporation. The system augments simple keyword searching with the beta version of its visual/texture based search engine. Currently, the website allows users to search within thumbnails provided from its image database.

Algorithm and methodology

The algorithm utilized in our testing of color-histogram approach to content based image retrieval is based on the paper written by Jain and Vailaya. The following is an outline of the method.

1. Read images in database and extract *YCrCb* format pixel information from images.
2. Create 48 bin normalized histograms for each of the *YCrCb* components of each image read from database. Thus, each image will have 3 histograms associated with it.
3. Read in a query image and extract *YCrCb* format pixel information
4. Create histograms for each of the *YCrCb* components of the query image.
5. Compute a Euclidean distance by comparing the query image histograms to that of each image in the database.
6. Sort images in database in order of ascending Euclidean distance to query image and return as result.

Extraction of *YCrCb* information

ISSN:2319-7900

The algorithm was implemented in Java and, thus, the built-in methods provided by Java's image class were utilized to retrieve an array containing pixel values in *YCrCb* format. As a result, only image formats compatible with Java's built-in methods were utilized. These consist of the most common formats including, JPEG, BMP, GIF and PNG.

Comparison

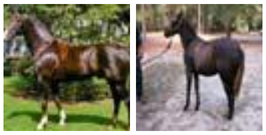
Once the histograms have been created, Euclidean distances are calculated. Differences are calculated for each bin by comparing the proportion of pixels of a certain intensity level in each level and then these differences are squared. The squared distances are summed together. The square root of this value is taken. This process is carried out for each histogram after which the average of the three values is taken.

Image collection and experiment set-up

All images used in our experiments are available online at: http://www.students.bucknell.edu/rc036/csci378/ the database utilized during program implementation underwent different stages. The preliminary database Consisted of images of comic superheroes due to their easily identifiable color schemes. Also, the initial database contained images that were used to the resilience of the image to transformations. Thus, images were taken and put through rotations, flips, resizing, brightening and darkening. The next stage database was expanded to include five different categories of images collected from various places on the internet as well as my personal collection of images: animals, colors, landscapes, structures, and superheroes Thus, in total there were 6 different categories: Lions, Flowers, Orchids, Horses, Aircrafts and Snowboarding. It is important to note that some of these pictures were only taken as thumbnails, thus, the results do not exactly match those found on the online services.

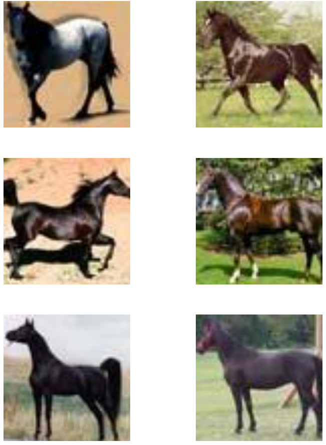
Result Analysis:

SAMPLE DATA BASE IMAGES:



Query Image

Retrieved Images



Conclusions:

We demonstrated by the various implementations of content based image retrieval systems, color histogram based comparisons can be easily combined using weights with techniques that extract other information from the image. Thus, this easy to implement technique of comparing images is an effective tool for accurate content based image retrieval.

References

[1] *AIRS - Advanced Image Retrieval Service*. http://www.imageclick.com/airs/sub/aboutAIRs.html (accessed October 2008).

[2] *Content-based image retrieval - Wikipedia*. 5 November 2008. http://en.wikipedia.org/wiki/CBIR (accessed November 8, 2008).

[3] Iqbal, Qasim. *CIRES: Content Based Image Retrieval System.* August 2007. http://amazon.ece.utexas.edu/~qasim/research.htm (accessed November 2008).

[4] Jain, Anil K., and Aditya Vailaya. *Image Retrieval Using Color and Shape*. Great Britain: Elsevier Science Ltd, 1995.

Biographies

FIRST A. AUTHOR received the B.E. degree in Computer Science and Engineering from the University of Madras, Chennai, TamilNadu, in 2003, the M.Tech. degree in Computer Science and Engineering from Bharath University, Chennai,TamilNadu, in 2012, respectively. Currently, He is an Assistant Professor, Dept. of Computer Science Engineering, Bharath University, Chennai.