Available online at www.sciencedirect.com
ScienceDirect
Displays 29 (2008) 260–267
DISPLAYS
www.elsevier.com/locate/displa

Empirical comparison of image retrieval color similarity methods with human judgment

Hock Chuan Chan *

National University of Singapore, 3 Science Drive 2, Singapore 117543, Singapore

Received 8 May 2007; accepted 16 August 2007
Available online 23 August 2007

Abstract

It is important to verify assumptions and methods of image retrieval against actual human behavior. A study was conducted to compare similarity methods of color histograms against human assessment of similarity. The similarity methods tested include basic histogram intersection, center histogram matching, locality histogram matching, and size-weighted histogram matching. 161 subjects participated in the empirical study. The findings, based on Spearman correlation analysis, showed that both the basic histogram intersection method and size-weighted histogram are very close to human assessment of similarity (Spearman correlation coefficient of 0.915). The other two are not close to human judgment on similarity. This study illustrates an alternative approach to evaluating matching algorithms. Unlike the usual measures of recall and precision, this approach emphasizes human validation. Fewer images are required with the use of statistical testing.

Keywords: Image retrieval; User issues; Histogram matching; Human similarity judgment

1. Introduction

With the growth of multimedia systems and databases, image retrieval remains as an important research area. Retrieval based on content is the dominant approach, compared with textual description and search [1]. For non-semantic information, color is probably the easiest for humans and computer systems to identify and describe. "The majority of retrieval techniques implement color histograms for image retrieval using color" [2]. Textures, shapes, and spatial relations, although less common, are also widely researched [3–5]. Thus, image retrieval by color features has been dominant [6–15]. Many algorithms have been proposed on how to extract color information from images, and how to match such color information from different images. "Similarity-based retrieval of images is an important task in many image database applications" [16, p. 115].

Typically, algorithms have been evaluated using recall and precision measures [1,12,17,18], or modified recall and precision measures [7]. Recall is the ratio of the number of relevant images retrieved to the total number of relevant images in the database. Precision is the ratio of the number of relevant images retrieved to the total number of images retrieved. These ratios can be manipulated by system parameters. For example, recall can be easily increased by retrieving more images, but at the expense of precision. Jose et al. advocates that more measures should be considered [19]. Recall and precision measure only one dimension of system performances. Other measures from the user perspectives, such as usability, and user acceptance, should also be considered.

Similarly, Nishiyama et al. state, "The algorithms of image retrieval operations have to suit user's subjective viewpoint, such as a similarity measure" [20, p. 30]. In the image retrieval area, there are very few empirical studies that compare computer algorithms against human

* Tel.: +65 6516 3393; fax: +65 6779 1610.
E-mail address: chanhc@comp.nus.edu.sg

0141-9382/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.displa.2007.08.013

judgment [21]. Tamura et al. studied humans on textual features [22]. Scassellati et al. conducted an experiment to compare human judgment and algorithms on shape similarity [23]. Nishiyama et al. studied 20 subjects and 3 target images using a particular image retrieval system, and found that a sizeable number (up to 35%) of them were unable to remember the location of objects in the images [20]. Gudivada and Raghavan conducted a study of spatial relation similarity algorithms against an expert's judgment [16]. It used 24 pictures with shapes in various relations with each other. Chan and Wang conducted an empirical study on human's ability to specify color percentages [24]. Jose et al. conducted an experiment with 8 subjects, comparing a textual search system with a system that allows visual spatial image specification [19]. They found statistically significant advantages for the later system. This supports the assumption popular among image retrieval researchers that textual retrieval is not as good as other visual retrieval methods. Payne et al. also compared textural algorithms against human measures of similarity [25]. Mojsilovic et al. conducted a comparison study using 28 human subjects and 25 patterns [26]. They discovered that humans used color and directionality in their similarity judgment.

To add on to the slowly growing body of knowledge on empirical validation with human ability and perception, this study compares image matching methods (similarity measures) against human matching. "Measuring the dissimilarity between images and parts of images is of central importance for low-level computer vision" [27, p. 1165].

Section two describes the matching methods used in this study. Section 3 describes the details of the empirical study. The data are analyzed in Section 4. Section 5 provides more discussions of the findings and the conclusion.

2. Matching methods used in this study

2.1. Simple histogram intersection matching

Each image in the database is represented using three primaries of a color space. The most common color space used is RGB. All the methods in this study use only RGB. Other color spaces, such as Munsell, CIELuv or CIELab, are not included [28–30]. Let the three primary colors be divided into k, l, m intervals, respectively. The total number of discrete color combinations (called bins) n is equal to $k \times l \times m$. Dividing each primary into 16 intervals will give a total of 4096 bins. A color histogram $H(M)$ is a vector $(h_1, h_2, \ldots, h_n)$, where each element h_j represents the number of pixels falling in bin j in the image M. These histograms are the feature vectors (indexes) stored as the index for retrieval searching. For example, Pass et al. utilized the RGB color space and quantized uniformly into 64 bins for their image retrieval system [31].

For retrieval, a query histogram constructed by the system, with input from the user, is matched against the histograms in the database. Typically, the user can specify the histogram directly, or through selecting or drawing a sample image. Some metric has to be used to estimate the "distance" or "similarity" between two histograms. This metric is then used as a criterion for deciding whether to retrieve an image for a given query. For example, the system can have a threshold value, i.e. only images closer than a predefined threshold value will be retrieved. Another common approach is to retrieve the closest predefined number of images.

Swain and Ballard proposed the histogram intersection metric based on the vector representation of the histogram [30]. Suppose I and J are the histograms of query image and database image, respectively, each containing N bins. The intersection (In) is defined as

$$\mathrm{In}(I,J) = \frac{\sum_{n=1}^{N} \mathrm{Min}(I_n, J_n)}{\sum_{n=1}^{N} I_n}.$$

In this metric, the intersection is incremented by the pixels common between the target image and the query image. The measure is finally divided by the total pixels in the query image as a normalization factor. This will generate a value between 0 and 1, where 0 is least similar, and 1 is most similar.

In this basic method, the whole picture is used to generate an image histogram, and matching is done by comparing bins equally. In the subsequent methods, slight variations will be introduced.

2.2. Center of image

Recently, many algorithms have been proposed to include richer information in a color histogram so that it will be more efficient in differentiating relevant and irrelevant images. For example, some people observe that images (usually photos) usually have the theme at the center, and the surrounding area, although contributing a large part in color components, usually do not directly relate to the theme. Including each area equally in the color histogram, as the basic color histogram algorithm does, simply ignores the different contribution of the two parts. Based on this assumption, many proposed that the surrounding part should be given lower weight when the two histograms are matched.

Stricker and Dimai observed that images are usually photos, and photographers almost always placed the object in the center, and concluded that the center is very important [32]. They proposed an approach that divided images into a center region (oval) and 4 surrounding regions, from which to extract color distribution features. When matching images, users can set weights to different region according to their importance, with the center region usually given higher weights than the surrounding regions. This approach, the authors claimed, significantly increases the discrimination power when compared with the basic color matching method.