Content Based Image Retrieval by Integration of Metadata Encoded Multimedia (Image and Text) Features

Mark Rorvig, Ki-Tai Jeong, Anup Pachlag, Ramprasad Anusuri, Sara Oyarce
Texas Center for Digital Knowledge, The University of North Texas
Denton, Texas 76203 USA

Abstract

Content Based Image Retrieval (CBIR) is the retrieval of imagery from a collection by means of internal feature measures of the information content of the images. In CBIR systems, text media is usually used only to retrieve exemplar images for further searching by image feature content. This paper describes a new method for integrating multimedia text and image content features to increase the retrieval performance of the system. The results indicate an increase in the precision of image retrieval by more than 300% due to the multimedia content feature integration process.

1 Introduction

At the WWW10 Conference in Hong Kong, two of us (*MR, KTJ*) released a poster paper describing a new method for image feature content metadata extraction and encoding <**http://www.www10.org.hk/cdrom/posters/p1142/index.htm**>. The paper described a novel approach to the extraction of image features and their representation as metadata tags in the header of the HTML record. A variation of this system was later used in the NIST TRECvid video retrieval trials where its precision score in the general search category was 0.59 <**http://trec.nist.gov/pubs/trec10/papers/UNT_final.pdf**>. Precision is defined as the ratio of relevant to non-relevant items retrieved in a search request.

The method of the WWW10 report had a number of advantages over conventional ones. First, it permitted expression of the image content features as XML/HTML header metadata content codes. Second, it supported a method of searching for images by simple subtraction of the values of an exemplar image from the stored values of the content codes for the collection images. Those images from the collection that differed from the exemplar image by less than a user controlled threshold were retrieved and the rest not retrieved.

In this paper a new technique is introduced for the derivation of text features as metadata integrated into the same content feature structure as values for image feature measures. The goal of this integration was to increase the precision of the image retrieval system. Image retrieval systems often use text to retrieve image exemplars for use in content based image feature searching. However, beyond that point, text usually remains unexploited. *In other words, the value of text in the content retrieval is lost due to the separation of the two processes of text search and image content search.*

The experiment we report uses the previously neglected value of the text information in the image retrieval process by encoding it in the same content feature structure as the image feature content. This technique permits both text and image information to be integrated in the search process by the simple feature subtraction method previously used for image content features alone. Competing methods for such integration typically use either expert rules or Bayesian evidence combination equations. (Please see Deb Roy's work at <**http://citeseer.nj.nec.com/309208.html**> for examples.)

2 Text Features

In the text feature technique, all unique words are extracted from a collection of image description documents. An image description document collection may thus be represented by creating a rectangular matrix of features in which rows are image document descriptions and the columns are individual words of text contained in those descriptions. For each cell of a row of features, the number of times a feature (a text word) is used in the image description document is recorded. The feature columns are then summed, and sorted by their sums from highest to lowest column sum values. This operation captures the frequency values of individual text terms and records it by its position in the row. These row values are further abstracted to binary values. Thus, if for any image description document there are values greater than zero in the columns of text terms sorted by frequency, they receive a "1", and otherwise a "0" in the column position. Image description document rows are thus transformed into binary strings.

To complete the transform to features, the binary strings of rows are divided into a set of convenient intervals of equal length. In this experiment, the more than 5000 individual columns of text features from 994 NASA images (the STS-82 Hubble Space Telescope Repair Mission) of the test collection are divided into twelve segments of equal length corresponding to the twelve image values used in the prior study (for a description of these measures please consult the WWW10 URL above). For each segment, a “code weight” is obtained by summing the number of binary “1” values within the segment. The weight of each code for a binary string segment of a row is then divided by the total sum of binary “1”’s for that row. This operation is repeated for all segments of all rows. In this manner, text terms of an individual image description document are reduced to a set of real numbers ranging in value from 0.0 to 1.0. These values are then recorded as metadata fields in the image description HTML records along with the values for image content features.

3 Procedure Description and Results

For this experiment, two parallel systems were built. The first <**http://archive4.lis.unt.edu/td/www**> system used only image content features. The second used both image and text features <**http://archive4.lis.unt.edu/tdt/www**>. From the NASA Image Archives, a set of 25 questions submitted specifically regarding STS-82 was collected from records of prior requests. The relevance judge for the results was the manager of the NASA Image Archives at the Johnson Space Center. This individual was asked for each question to consider all the presented images from both systems and mark the ones directly responsive to each question. Her responses were recorded, cross-referenced to the system of origin for each image, and scored. For the system without text features the precision score was 0.13. For the system using both image and text features, the precision score was 0.43. Integration of text feature codes into the image content code structure increased the precision of the image retrieval system by 331%.

ACKNOWLEGEMENTS

This study was supported by Intel Corporation, the Special Libraries Association, the National Aeronautics and Space Administration, and the Texas Center for Digital Knowledge. All correspondence should be addressed to Mark Rorvig <mrorvig@unt.edu>, 940-300-5344, (fax) 940-565-3101.