PRACTICAL MPEG-7 IMAGE INDEXING & RETRIEVAL FOR UNDERGRADUATES

P. Andoutsos, A. Kushki, A.N. Venetsanopoulos

Edward S. Rogers Sr. Department of Electrical and Computer Engineering
University of Toronto
10 King's College Road
Toronto, Ontario, Canada
M5S 3G4

ABSTRACT

A practical component to an upper year electrical and computer engineering course in information engineering or multimedia systems is presented. Developed to introduce students to image indexing and retrieval, this experiment used a Microsoft Windows binary of the freely available MPEG-7 *Experimentation Model* from ISO for image description. The experiment was broken into three interdependent stages spanning a three week period. Each stage demonstrated various concepts associated with image indexing and retrieval. Students required C/C++ programming skills to parse binary encoded MPEG-7 descriptions, output ASCII description data, and perform retrievals on indexed information. This laboratory was distributed in nature, and each participating student was made to work on a small, mutually exclusive image set. This experiment drives home the urgency of the media indexing problem in light of rapidly growing market need for multimedia management applications resulting from a huge influx of multimedia data. Through feedback, student views indicated that despite being long and time consuming, this experiment provided an interesting, original and stimulating experience.

1. INTRODUCTION

The International Standards Organization's (ISO) MPEG-1 and MPEG-2 standards are common in most upper-level undergraduate courses in multimedia. Yet, little emphasis is placed on state-of-the-art concepts and techniques such as MPEG-7. This is partly because of the abrupt change in paradigm from data representation to data description. In order to maintain a very up-to-date curriculum at the University of Toronto, a practical laboratory was developed to expose students to content-based indexing and retrieval using MPEG-7. In light of recent attempts in creating curricula which address engineering of multimedia and creative technologies[1], it can be said that typical electrical or computer engineering graduates lack the expertise necessary to satisfy the needs of industries that rely on multimedia applications. In fact, many employers actually find that the hiring of graduates from technical schools or colleges (where curricula are revised nearly every year) is often more effective than recruitment from universities (where technical curricula are typically updated every four to eight years)[2]. This paper presents a current, hands-on experiment which provides students with an up-to-date taste of ideas that are growing ever more necessary in the digital media realm. This last point is of critical importance because the recent multimedia data flood has necessitated the automatic indexing and retrieval of digital audiovisual assets. This experiment was broken up into three distinct phases over three weeks, each of which addressed one or two specific concepts. The first phase addressed ground truth extraction, the second required students to employ MPEG-7 XM binaries to assigned images for the generation of descriptor data, and the last phase required students to execute retrieval simulations, and a quantitative performance analysis with respect to the extracted ground truth.

The paper is structured as follows: Section 1.1 presents the preliminaries necessary for the proper flow of the experiment, while Sections 2.1 through 2.3 outline the three stages of the experiment. Closing remarks, student feedback responses, and possible changes to this experiment are presented in Section 3.

1.1. Preliminaries

To ensure proper flow of this experiment, a preliminary planning stage was required. It was the job of the laboratory supervisor to properly partition the class and appropriately assign descriptor responsibilities, ground truth images, and images for description. Figure 1 provides a block diagram showing the experiment's three component phases along with an explanation of student expectations, benefits, resources, etc. In the preliminary step, two mutually exclusive subsets of database images were distributed to participants; one set for ground truth extraction and retrieval, and a second set

for image description. While the size of the first set was determined by class size alone, the second set was also based on the number of descriptors used. This preliminary step was also necessary to ensure that students were aware of the descriptors for which they were responsible.

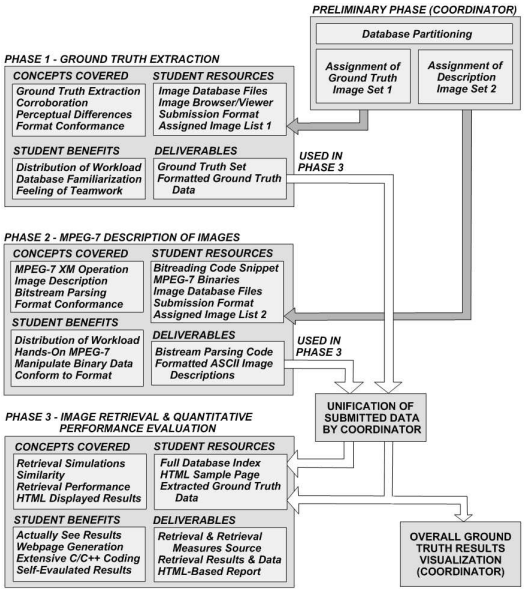

Fig. 1. Flowchart showing overall structure of the lab.

2. EXPERIMENT

2.1. Phase 1 - Ground Truth Extraction

For a general image database, the generation of ground truth data is a very laborious process. This is because all database images need to be compared to candidate images by multiple experts before final decisions on their relevance can be made. A simplified version of such an approach was used for the this experiment, and each student was required to take one image from their first assigned image set (about five images) and visually compare it to each database image. The goal during this phase was for students to generate a list of images which they felt were 'similar' to each of their assigned images. A corroboration step was performed subsequent to this in attempts to try and eliminate some of the perceptual differences which exist across viewers. This was performed in small student groups where the images of candidate ground truth sets were checked by two peers using a best two out of three approach. It should be mentioned that ground truth data set for each student had no relevance to other class members because of their mutually exclusive nature. This allowed students the freedom of open co-operation, and ensured that students did their own work.

2.2. Phase 2 - MPEG-7 Image Description

In the second phase of this laboratory the focus was shifted to the application of the MPEG-7 eXperimentation Model (XM) for representing images according to ISO standardized descriptions. Three complimentary descriptors were selected due to their simplicity, and straightforward bitstream encodings. These descriptors were the *Dominant Color*, *Edge Histogram*, and *Color Structure*[3] descriptors. Each student was responsible for applying the MPEG-7 XM to their second set of assigned images using one of three descriptors. Since the application of the XM resulted in a single bitstream encoded according to the ISO MPEG-7 standard, parsing was necessary to produce easily manipulated and understood ASCII text data. Students were required to write C/C++ code which read the source bitstream file along with the associated image list from disk. The parser had to properly read binary MPEG-7 data and output ASCII text along with associated image filenames according to a specified format. Use of a predefined submission format was necessary to simplify the coordinator's task of collecting of and unifying student submitted descriptor data. The three unified indexes (one per descriptor) were used in phase three for retrieval. The flowchart in Figure 2 illustrates phase two's processing of assigned images by student 'n' who subsequently performs bitstream parsing and formatting of descriptor data before submitting it to the lab coordinator. Upon collection of all student submissions, the lab coordinator unifies all all student descriptor data to create amalgamated ASCII index files for each descriptor.

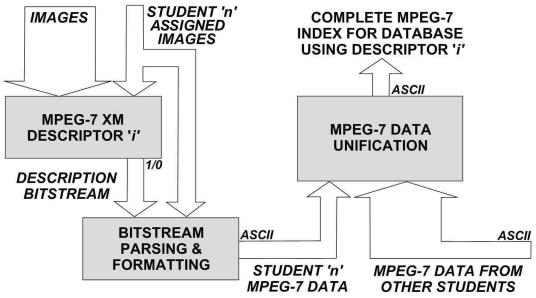

Fig. 2. Flowchart for phase two. Each student was responsible for applying the MPEG-7 XM on an assigned set of images for a single descriptor. The ASCII descriptor data was collected and unified by the coordinator into a larger index file to be used by students in phase three.

2.2.1. Descriptors Employed

The Dominant Color descriptor provides a simple, yet compact representation of the most important colors in a particular image. The dominant color description vector $\overrightarrow{d_{DC}} = \{(p_i, R_i, G_i, B_i), i = 0, 1, ..., N_{DC}\}$ consists of RGB color

values R_i, G_i, B_i, along with their respective percentages p_i. N_{DC} is the total number of dominant colors in the image and $\sum_{i=1}^{N} p_i = 1$. To capture color information invisible to the Dominant Color descriptor, the Color Structure descriptor was used. This descriptor provides an indication of the degree of connectedness of pixels of a particular color (i.e. whether they exist in a single, connected blob or distributed across the image in smaller, unconnected regions). The Color Structure descriptor is represented by a single vector $\overrightarrow{d_{CS}}$ consisting of N_{CS} integer values spanning the range $[0, 255]$. In this case, N_{CS} was chosen to equal 128 from the set $\{256, 128, 64, 32\}$. This is unlike $\overrightarrow{d_{DC}}$ which has variable length depending on the number of dominant colors present in the image. Lastly, the Edge Histogram texture descriptor suggested by MPEG-7 was selected to enrich the descriptor set because of its simplicity. Typically, texture analysis and description algorithms are difficult to grasp because of the highly specialized mathematics which are required (e.g. The *Homogeneous Texture* descriptor in the MPEG-7 descriptor set employs Gabor functions in angular and radial directions). The 80-dimensional vector $\overrightarrow{d_{EH}}$ of quantized values representing the Edge Histogram, provided a very straighforward description of directional edge histograms in in the directions 0^o, 45^o, 90^o and 135^o for sixteen subimages [4].

2.3. Phase 3 - Retrieval and Analysis

The last phase of the experiment consisted of three components and required the combination of data and results from phases one and two. The first component used the amalgamated descriptors created from phase two's unified indexes to facilitate image retrievals. Using appropriate distance measures, and MPEG-7 descriptor data, retrieval simulations were performed. The second component dealt with the analysis of the retrieval results using quantitative retrieval measures along with ground truth data extracted during phase one. The distance measures used to gauge similarity are outlined in Section 2.3.1, while an explanation of the retrievals and analyses performed is provided in Section 2.3.2.

2.3.1. Distance Measures & Similarity

Three different distance measures for similarity calculation were used here. Since each measure is defined on its own interval, no hybrid queries across descriptors were performed to maintain simplicity. First,distances between Color Structure descriptors, were calculated using the the *Euclidean Distance* ($L2$-norm). In the expression in (1), the overall distance $D_{CS}(A, B)$ between images A and B is based on a sum of the values in the 128-dimensional data vector.

$$D_{CS}(A, B) = \sqrt{\sum_{0}^{127} \left[\overrightarrow{d_{CS_i}}(A) - \overrightarrow{d_{CS_i}}(B)\right]^2}. \quad (1)$$

For Edge Histogram data, similarity values were generated using the $L1$-norm (*Manhattan Distance*) shown in (2) where $D_{EH}(A, B)$ represents the overall distance between images A and B, and $d_{EH,i}(I)$ denotes the i^{th} element of the 80-dimensional edge histogram data vector for image I.

$$D_{EH}(A, B) = \sum_{i=0}^{79} \left|\overrightarrow{d_{EH,i}}(A) - \overrightarrow{d_{EH,i}}(B)\right|. \quad (2)$$

Unlike the fixed length data vectors $\overrightarrow{d_{CS}}$ and $\overrightarrow{d_{EH}}$, a much more rigorous distance measure was necessary to accommodate the image-dependent vector length of the Dominant Color descriptor. As outlined in [5], a distance measure incorporating the $L2$-norm between RGB color values, color coverage values and a distance threshold was employed to compare Dominant Color descriptor data. For two images A and B, each respectively containing N_A and N_B dominant colors, the overall distance between the description vectors $\overrightarrow{d_{DC}}(A)$ and $\overrightarrow{d_{DC}}(B)$ is:

$$D^2_{DC}(A, B) = \begin{array}{c}(\sum_{i=0}^{N_A-1} p^2_{A_i} + \sum_{j=0}^{N_B-1} p^2_{B_j} - \\ 2\sum_{i=0}^{N_A-1}\sum_{j=0}^{N_B-1} a_{(i,j)} p_{A_i} p_{B_j})\end{array} \quad (3)$$

In (3), the coefficient $a_{(i,j)}$ is used for cross color similarity between $\overrightarrow{d_{DC}}(A)$ and $\overrightarrow{d_{DC}}(B)$. This approach is necessary since N_A is typically not equal to N_B. Each coefficient was calculated as in [5] where color distances are determined by employing the L2-norm between RGB color vectors, and the values d_{max} and T_d explained therein were set to 30 and 20 respectively.

2.3.2. Retrieval & Analysis

The first component of phase three, required students to make retrieval queries for each of the images in their first assigned image set. This resulted in a group of ranked retrieval results which were formatted using HTML. Code for performing retrievals was implemented using C/C++, and student binaries were required to a) read in amalgamated descriptor files, b) execute retrievals from user queries using an assigned descriptor with its corresponding distance measure, and c) generate HTML webpages of retrieval results. HTML output was chosen for visual inspection of results because of its flexibility and simplicity in formatting and displaying images (jpg format), as well as because of its ubiquity. A sample student retrieval is shown in Figure 3.

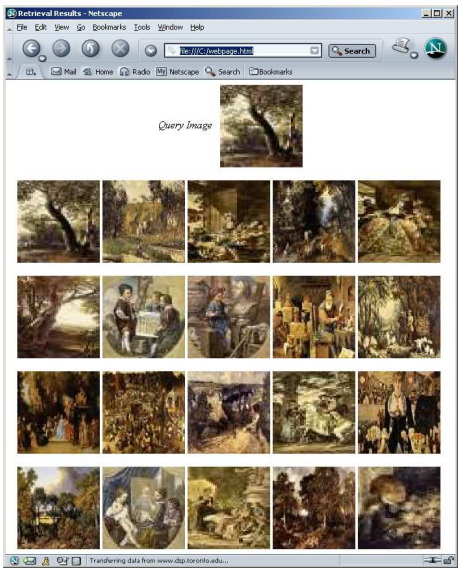

Fig. 3. Sample student submission of retrieval results using a conventional web browser.

The second component of phase three involved the calculation of retrieval effectiveness using the ranked retrieval results, and the ground truth extracted in phase one. The quantitative retrieval evaluation measures used here were the same used by the ISO standardization committee for query-by-example, and included the Average Rank (AR), Modified Retrieval Rank (MRR), Normalized MRR (NMRR) and Average NMRR (ANMRR)[6]. All calculations were done by students manually, and effectiveness values were indicated on their webpages of retrieval results which eventually became their formal reports for the experiment.

3. CONCLUSIONS & FUTURE IMPROVEMENTS

The paper presented here outlined a practical experiment for an upper-level undergraduate course in multimedia systems in electrical and computer engineering or computer science. The development of this hands-on experiment was inspired by the lack of up-to-date laboratories which are typically offered in many post-secondary engineering schools. The laboratory presented was broken up into three distinct phases, each of which addressed specific concepts related to the content-based retrieval of images. The experiment was performed using a distributed approach because of the large sizes of image databases typically used, and assigned to students small subsets of the database for which they were responsible. From questionnaires, student feedback about this experiment was collected. Some of the responses were:

- "This is the one of the most interesting labs I have done in the 4 years I have been here."
- "An extremely long lab."
- "Phase 1 was good, so students will be able to get a general idea of how difficult it is to extract images in an image database...Phase 2 was a bit tedious, but we learned how to parse the bitstream...Phase 3 was informative, but long. Overall, though, it was a good learning experience."
- "I think this lab is very interesting. Even though my final results did not match my ground truth, I was able to see why it failed and get a better understanding of how hard it is to extract information from images."
- "I think this lab can be improved if the ground truth extraction was done with more 'quality control'. By this I mean instead of assigning five different images to everyone, have the entire class do the same five images. You can set the minimum number of ground truth to a high number like 30-40, and then get the 'real' ground truth by only keeping the intersection."

The overwhelming consensus of the returned questionnaires was that this experiment was very long and problematic in ground truth extraction. At the same time however, it was both fresh, and interesting. A number of variations are planned for this experiment in the future. First, extending the lab to cover four weeks will be attempted to try and split up the two components in phase three. As suggested by student feedback, the ground truth stage needs to be altered. To this end, the assignment of a specific set of ground truth/query images will be considered. These two changes will definitely alter the face of the experiment. Thus, turning the lab into a competitive project will be pursued to have students vie to achieve the best retrieval performance.

4. REFERENCES

[1] J. M. Mendel, "Establishing academic programs in integrated media systems", *IEEE Signal Processing Magazine*, vol. 16, no. 1, pp. 67-76, January, 1999.

[2] A. E. Paton, "What industry needs from universities for engineering continuing education", IEEE Transactions on Education, Vol. 45, no. 1, pp 7-9, February. 2002.

[3] ISO/IEC 15938-3:2001, "Multimedia Content Description Interface - Part 3: Visual", Version 1.

[4] D. K. Park, Y. S. Jeon, C. S. Won, S.-J. Park, "Efficient use of local edge histogram descriptor,", *Proceedings of ACM International Workshop on Standards, Interoperability and Practices*, Marina del Rey, California, USA, 2000, pp. 52-54.

[5] Y. Deng, B. S. Manjunath, C. Kenney, M. S. Moore, H. Shin, "An efficient color representation for image retrieval", *IEEE Transactions on Image Processing*, vol. 18, no. 1, Jan. 2001.

[6] B. S. Manjunath, P. Salembier, T. Sikora, "Introduction to MPEG-7", Wiley, 2002.