BILKENT UNIVERSITY MULTIMEDIA DATABASE GROUP AT TRECVID 2008

Onur Kucuktunc, Muhammet Bastan, Ugur Gudukbay, Ozgur Ulusoy

Department of Computer Engineering, Bilkent University, Ankara, Turkey
{onurk, bastan, gudukbay, oulusoy}@cs.bilkent.edu.tr
http://www.cs.bilkent.edu.tr/~bilmdg

ABSTRACT

Bilkent University Multimedia Database Group (BILMDG) participated in two tasks at TRECVID 2008: content-based copy detection (CBCD) and high-level feature extraction (FE). Mostly MPEG-7 [1] visual features, which are also used as low-level features in our MPEG-7 compliant video database management system, are extracted for these tasks. This paper discusses our approaches in each task.

1. INTRODUCTION

TRECVID have become an indispensible evaluation for research groups working on content-based analysis of and retrieval from digital video for a couple of years [8]. In 2008, TRECVID introduces two real-world situations as new task evaluations: surveillance event detection, and content-based copy detection.

Multimedia Database Group in Bilkent University participated in TRECVID for the first time this year. As a new and small team, we were able to submit runs for CBCD and FE tasks as outputs of our baseline systems. In this paper we present our methods and evaluation results for each task.

Rest of the paper is organized as follows: Section 2 represents visual descriptors used in preprocessing stage. Section 3 addresses our work on content-based copy detection task. In Section 4, detailed information of our high-level feature extraction system is discussed. Finally, the evaluations and conclusions are given in Section 5.

2. DATA PREPROCESSING

2.1 Overview

For the tasks we participated in TRECVID 2008, a single pre-processing on video data is performed. This stage requires the extraction of visual information from selected video frames.

We used FFMPEG [2] library to decode the mpeg videos, and OpenCV [3] library for image manipulation. Shot boundary data is provided by NIST. An MPEG-7 feature extraction library adapted from MPEG-7 XM software is used for extracting MPEG-7 visual features [4]. eXperimentation Model (XM) is the reference software, which implements all the reference code of MPEG-7 standards [1]. We are only interested in descriptors and extraction tools; thus, the extractor parts are separated from the framework and combined together in order to extract the visual features from the given frames in a parametric manner.

2.2 Visual Descriptors

A set of visual features defined in MPEG-7 standards is selected. From color descriptors: scalable color (SCD), color layout (CLD), and color-structure (CSD); from texture descriptors, homogenous texture (HTD), and edge histogram descriptors (EHD) are chosen. Definitions of these visual descriptors are presented in detail:

2.2.1 Color Descriptors

Scalable Color Descriptor: SCD is a color histogram in HSV color space, and encoded by a Haar transform. It is generally used in image retrieval systems based on color feature.

Color Layout Descriptor: CLD essentially captures the layout information of color feature. Because of its high retrieval efficiency and small computational costs, CLD is preferred in image and sequence matching and sketch queries.

Color-Structure Descriptor: Color structure of an image holds both the color content (like a color histogram) and also the structure of this content. CSD provides a better retrieval performance on natural images compared to ordinary color histograms

2.2.1 Texture Descriptors

Homogenous Texture Descriptor: HTD provides a precise quantitative description of a texture that can be used for accurate search and retrieval. The computation of this descriptor is based on filtering using scale and orientation selective kernels.

Edge Histogram Descriptor: EHD represents the spatial distribution of four directional edges and one non-directional edge. It provides better performances on image matching with non-uniform edge distribution.

3. CONTENT-BASED COPY DETECTION

3.1 Task Description

Content-based copy detection task is defined as detecting copies of a video derived by various transformations such as addition, deletion, modification (of aspect, color, contrast, encoding), camcording, etc. Growing broadcasting of video content on TV channels, video blogs, and on other media sources makes copy detection an even more challenging problem.

For CBCD task in TRECVID 2008, participating groups are required to determine the location of each query, if any, with a decision score within a test collection of approximately 200 hours of Sound & Vision data.

Each query is constructed by taking a segment of varying length from both reference dataset and some videos not in the database (to test false positive), and then applying one or more transformations to this query segment. Final list of transformations includes 10 modification types: camcording, picture in picture, insertion of pattern, strong reencoding, change of gamma, decrease in quality, and combinations of some or all of modifications.

3.2 Method

Our method for CBCD is very similar to the technique used in content-based image retrieval systems. Firstly, signature of a query, extracted by MPEG-7 descriptors mentioned previously in data preprocessing section, is generated. Then, this signature is compared with the signatures of all the selected and preprocessed frames of video dataset. If the computed similarity is above a threshold, we consider this match as a copy in our system.

In more detail, signatures of every first frame per 2 seconds in a query video are compared to the center frames of every shot in the dataset. We have used appropriate similarity measure (discussed in 3.3) and a variable-weighted feature similarity calculation technique (3.4) in our method.

3.3 Similarity Measure

Similarity measures for different MPEG-7 descriptors are explored and compared by Eidenberger in [5]. Based on his research, we selected Meehl index, pattern difference, and city block distance (L_1) for similarity calculations.

3.4 Variable-weighted Feature Similarity Calculation

Use of single visual feature for similarity calculations for such task is absolutely insufficient because of different types of transformations. Similarly, assigning constant weights to each feature does not work correctly most of the time, since different features are more discriminative than others for different query images or transformation types.

In our system, we used a new technique based on weighting a feature regarding its success rate. Lowe's SIFT matching algorithm [6] inspired us to develop this method. Success rate of a feature is defined as its discriminative property, calculated as the ratio of similarity values of the most similar match to the 5th one. Ratio of the most similar one to the second is not preferred because there may be visually similar shots in the same video.

3.5 Sample Results

Here we present some of the successful detections; even though complex transformations are used.

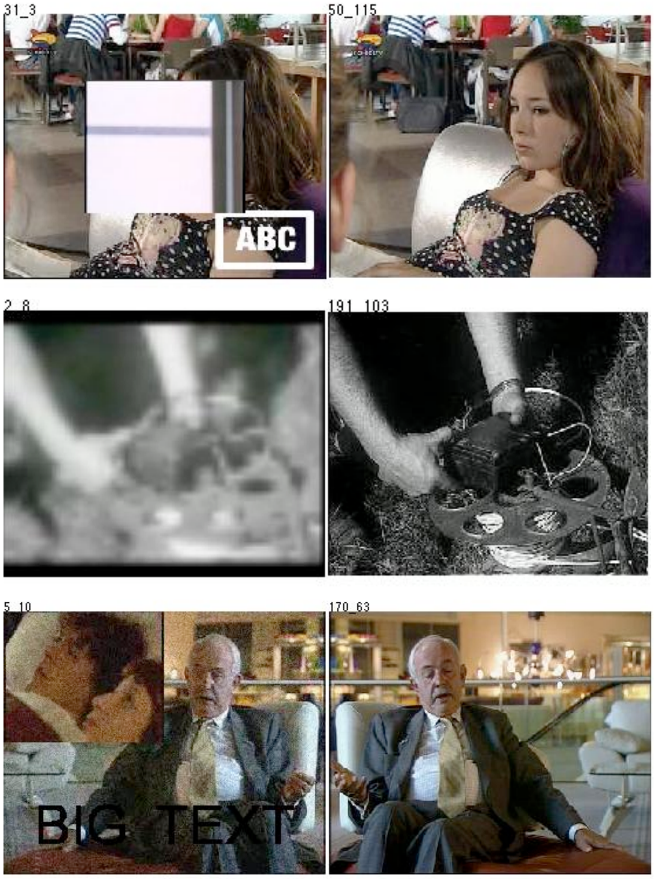

Figure 1. Frames in queries (left), and the detected copies (right)

4. HIGH-LEVEL FEATURE EXTRACTION

4.1 Task Description

The high level feature extraction task is defined as follows: given the feature test collection, the common shot boundary reference for the feature extraction test collection, and the list of feature definitions, the system will return for each feature the list of at most 2000 shots from the test collection, ranked according to the highest possibility of detecting the presence of the feature. Each feature is assumed to be binary, i.e., it is either present or absent in the given reference shot.

4.2 Method

Considering all 20 high-level concepts to be detected, it is necessary to utilize the multi-modal information present in videos to achieve a good detection performance. In addition to color and texture features, for instance, audio features can be used as primary or secondary information to detect the concepts *Singing* and *Domonstaration_or_Protest*, while motion information can be incorporated for concept *Airplane_flying*. Moreover, information related to the spatial relations between regions (e.g., *Sky* and *Airplane_flying*, *Sea* and *Boat_Ship*) should contribute to the overall performance. Finally, a hierarchical classification approach would probably be a good starting point to discriminate between, for example, *Indoor* (*Classroom*, *Kitchen*…) and *Outdoor* (*Street*, *Harbor*, *Cityscape*…).

In our submission, we were only able to complete a baseline system based on one-class classification using nearest neighbors. Hence, for each concept a separate classifier is designed.

We participated in the collaborative annotation effort organized by LIG [7] to annotate the TRECVID 2008 development data for the 20 high-level concepts and used the final annotations in our system development.

- **Training Set.** For each concept, we have selected a small set of frames that best represents that concept by examining the positive annotations.
- **Keyframe Selection.** The FE task requires us to find all the shots containing a given concept. The concept may be present only in a small part of the shot; therefore, representing a shot with a single keyframe is not adequate. We, therefore, selected several frames for each shot depending on the amount of change in color and edge descriptions of the frames.
- **Features.** We have used MPEG-7 color (SCD, CSD, CLD, DCD) and texture (HTD, EHD) descriptors with distance metrics suggested by MPEG-7.
- **One-class classification.** For scene related concepts we designed a simple one-class classifier using a selected subset of MPEG-7 color and texture features. For each frame in test set, its average distance to the nearest k positive instances of a concept is computed. If the distance is below a predetermined threshold, the shot containing this frame is considered as positive for that concept. Finally, positives are ranked according to their distances and presented to the user.
- **Object Detection.** Some concepts require specifically designed object detectors for better performance (e.g., *Hand*, *Telephone*, etc.). As an example, for concept *Two_People*, we employed a face detector (OpenCV) to detect and count the number of faces in each frame. Then, we ranked the shots according to occurrence of two faces within each shot, giving us a baseline detector for the concept, which can only detect two people whose faces are visible and detectable by our detector.

5. CONCLUSIONS

In our first participation in TRECVID, we submitted outputs of our baseline systems for CBCD and FE tasks utilizing mostly the MPEG-7 color and texture descriptors. We are improving upon our baseline systems for the future TRECVID contests.

6. REFERENCES

[1] Martinez JM (2001) Overview of The MPEG-7 Standard. ISO/IEC JTC1/SC29/WG11 N4031.

[2] OpenCV, Open Conputer Vision Library, http://sourceforge.net/projects/opencvlibrary.

[3] FFMPEG, http://ffmpeg.mplayerhq.hu.

[4] MPEG-7 XM Software, Institute for Integrated Circuits,Technische Universität Munchen, Germany, June 2001. http://www.lis.e-technik.tu-muenchen.de/research/bv/topics/mmdb/e_mpeg7.html.

[5] Eidenberger, H., Distance measures for MPEG-7-based retrieval, 5th ACM SIGMM international Workshop on Multimedia information Retrieval (MIR '03), ACM, New York, NY, 130-137.

[6] Demo Software: SIFT Keypoint Detector, http://www.cs.ubc.ca/~lowe/keypoints/

[7] Stéphane Ayache, Georges Quénot, Video Corpus Annotation Using Active Learning, 30h European Conference on Information Retrieval (ECIR'08), pp 187-198, Glasgow, March 30-April 3, 2008.

[8] Smeaton, A. F., Over, P., and Kraaij, W., Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06.