2009 Eighth International Symposium on Natural Language Processing

A Content-based Information Retrieval System for Video Searching

José M. Perea-Ortega, M. Teresa Martín-Valdivia, Arturo Montejo-Ráez and L.Alfonso Ureña-López

Abstract— **This paper describes a basic architecture for retrieving images previously extracted from video files. Our approach is made up of two main subsystems: the speech-based retrieval module and the image-based retrieval module. The aim of the experiments presented in this work is to establish a baseline approach to resolve the automatic video image retrieval task, making use of the speech content transcripts and the key frames extracted from video files. The main conclusion indicates that the use of fusion strategies by merging the text and visual data of the queries works better than those approaches that use only the textual part or visual part of the queries separately. Nevertheless, the results obtained confirm that in a content-based IR system it is more desirable to give more weight to the documents retrieved by the speech-based IR subsystem than those retrieved by the image-based IR subsystem.**

I. Introduction

MULTIMEDIA content retrieval is a challenging research field that has drawn significant attention in the multimedia research community. With the rapid growth of multimedia data, methods for effective indexing and search of visual content are decisive. Search in non-textual unstructured content, such as image and video data, is not yet effective. A common approach for video retrieval is to apply conventional text search techniques to the associated speech transcript. This approach works fairly well for retrieving named entities, such as specific people, objects, or places. However, it does not work well for generic queries related to general settings, events or people actions, as the speech transcript rarely describes the background setting or the visual appearance of the subject. The most substantial work in this field is represented by the TREC Video Retrieval Evaluation[1] (TRECVID) community, which focuses its efforts on evaluating video retrieval approaches by providing common video data sets and a standard set of queries.

In the automatic video image search task the system takes queries as input and produces results without any human intervention. Given a search test collection of video files, a multimedia statement of information need (user query), and a common key frame boundary reference for the search test collection, the system is expected to return a ranked list of video images or key frames from the test collection, which best satisfy the query [1].

The user queries can be composed of a textual requirement and a set of sample images and/or videos. The data collection consists of sound and video files, a master shot or key frame boundary reference and a Automatic Speech Retrieval and Machine Translation (ASR/MT) output [2]. All these resources are explained in the Section 3.

In this paper, we present a basic system developed for video image searching based on the speech content transcript from video data. In the field of image retrieval we have participated in the ImageCLEF campaign for four years [3]-[8], in TRECVID 2007 [9] and VideoCLEF 2008 [10], [11].

The rest of the paper is organized as follows. The system overview is explained in the Section 2. Resources and data used for the experiments are described in Section 3. Then, the experiments and results are shown in Section 4 and finally, the conclusions are discussed in Section 5.

II. System Overview

The architecture of our automatic video search system is a combination of text-based retrieval and image-based retrieval. The experimental study presented has been based on the textual information available from ASR/MT output. In the image retrieval any visual feature from images has been considered and we only use a content-based image retrieval tool like GIFT (The GNU Image-Finding Tool, http://www.gnu.org/software/gift/). Finally, a fusion process which merges the results from speech-based retrieval and image-based retrieval is applied, following several procedures. Therefore, our approach is composed of three subsystem: the speech-based retrieval module, the image-based retrieval module and the fusion process. Fig. 1 shows the overview of the automatic video search system proposed.

A. Speech-based retrieval

For the speech-based retrieval we have generated a XML document per scene, using an information retrieval (IR) system for indexing and searching them. Lemur (http://www.lemurproject.org/) has been used as IR system.

For generating the XML documents, it has been necessary to segment the speech transcriptions in *text per scene*. We

Manuscript received June 15, 2009. This paper has been partially supported by a grant from the Spanish Government, project TIMOM (TIN2006-15265-C06-03), project RFC/PP2008/UJA-08-16-14 and project FFIEXP06-TU2301-2007/000024 granted by the Andalusian Government. We would like to thank the TRECVID organization in general and Paul Over in particular.

José M. Perea-Ortega is with the Computer Science Department, University of Jaén, E-23071 Spain (phone: +34-953-211956; fax: +34-953-212472; e-mail: jmperea@ujaen.es).

M. Teresa Martín-Valdivia is with the Computer Science Department, University of Jaén, E-23071 Spain (e-mail: maite@ujaen.es).

Arturo Montejo-Ráez is with the Computer Science Department, University of Jaén, E-23071 Spain (e-mail: amontejo@ujaen.es).

L. Alfonso Ureña-López is with the Computer Science Department, University of Jaén, E-23071 Spain (e-mail: laurena@ujaen.es).

[1] http://www-nlpir.nist.gov/projects/trecvid/

978-1-4244-4137-2/09/$25.00 ©2009 IEEE