SINAI at TRECVID 2007

Manuel C. Díaz-Galiano, José M. Perea-Ortega, María T. Martín-Valdivia,
Arturo Montejo-Ráez, L. Alfonso Ureña-López
Department of Computer Science. University of Jaén, Jaén, Spain
{mcdiaz,jmperea,maite,amontejo,laurena}@ujaen.es

Abstract

This paper describes the first participation of the SINAI[1] group of the University of Jaén in TRECVID 2007. We have only participated in the automatic search task. Our approach is a very simple system made up of three main modules: the text-based retrieval subsystem, the image-based retrieval subsystem and the fusion module. We have submitted several runs exploring fusion of both textual and visual lists. Also the effect of text expansion in the topics has been of our concern:

- F_C_1_SINAI_1: Baseline run using only ASR/MT text features and topics without text expansion.
- F_C_2_SINAI_2: Baseline run using only ASR/MT text features and topics with text expansion.
- F_C_2_SINAI_3: Baseline run using only visual test data.
- F_C_2_SINAI_4: Run using text and visual features mixing the *text shots* recovered from text-based retrieval subsystem and *image shots* recovered from image-based retrieval subsystem, following the *RoundRobin*[7] algorithm.
- F_C_2_SINAI_5: Run using text and visual features mixing the lists from both subsystems, including the 75 percent of best *text shots* recovered by the text-based retrieval subsystem and the 25 percent of best *image shots* recovered by the image-based retrieval subsystem, with topics without text expansion.
- F_C_2_SINAI_6: Like the previous one but using topics with text expansion.

With these experiments we have tried to establish a baseline study of automatic search task of TRECVID, in order to improve the system in the future. The results of the runs submitted indicate that using visual data does not seem to improve the overall results compared to the text-only baseline. On the other hand, the expansion of the text topics with synonyms does not improve the baseline result either.

1 Introduction

This is the first year that the SINAI group from the University of Jaén participates in the search automatic task from TRECVID 2007. We have experience in the ImageCLEF campaign[3][10] for three years[6][1][2]. Our experimental study is based on three subsystems: the ASR/MT[4] (Automatic Speech Retrieval/Machine Translation) text retrieval subsystem, the image retrieval and the fusion modules. The fusion module mixes the results of both previous subsystems, following several procedures. We submitted a total of six fully automatic runs. These experiments were a combination of text features only (ASR/MT), visual features only and a mixture of both. All of our runs are of type C (i.e. trained other than types A and B), according to TRECVID

[1] http://sinai.ujaen.es/wiki/index.php/PortadaEn

guidelines[13], and do not make use of other data, such as the LSCOM[2] (Large Scale Concept Ontology for Multimedia) concepts.

The rest of the paper is organized as follows. The search task is described in Section 2. The system overview, experiments and results are explained in Section 3 and Section 4. Finally, the conclusions and future work are discussed in Section 5.

2 Search task

Search is a high-level task which includes at least query-based retrieval and browsing. The task is as follows: given the search test collection, a multimedia statement of information need (topic), and the common shot boundary reference for the search test collection, return a ranked list of at most 1000 common reference shots from the test collection, which best satisfy the need[13]. There are several modes allowed in the search task:

- Fully automatic. System takes query as input and produces result without any human intervention.
- Manually-assisted. Human formulates query based on topic and query interface and system produces result without further human intervention.
- Interactive. Human reformulates query based on topic, query, and/or results and system produces result without further human intervention on this invocation.

This year SINAI group has only participated in **fully automatic** search task, with a very simple system that it has not been trained with common TRECVID development collection data or another common annotations.

3 System overview

The overall architecture of our automatic search system is a combination of text-based retrieval and image-based retrieval. The experimental study presented has practically been based on the textual information available from ASR/MT output. In the visual retrieval we have not considered any visual features and we only use a content based image retrieval tool like GIFT (The GNU Image-Finding Tool)[3]. Figure 1 describes the overview of SINAI automatic search system.

3.1 Text-based retrieval

Two alternatives are approached for the text-based retrieval subtask:

- To generate a document by scene (shot) and to use an information retrieval system.
- To convert each video in a document where every scene in the video corresponds with a paragraph or passage, supported by an information retrieval system based on passages[5].

In our experiments we have chosen to generate a document per scene and to use LEMUR[4] like text-based retrieval system. For this, we have implemented a XML architecture that collects all related information for each video file. It has been necessary to segment the transcriptions of the speech recognizer in *text by scene*. We have considered a scene as a *shot* or *mastershot*.

In order to construct this architecture, two alternatives can be chosen:

1. To obtain the *video segments* of each video file and to find what *audio segments* belongs to each *video segment*.

[2] http://www.lscom.org/
[3] http://www.gnu.org/software/gift/
[4] http://www.lemurproject.org/

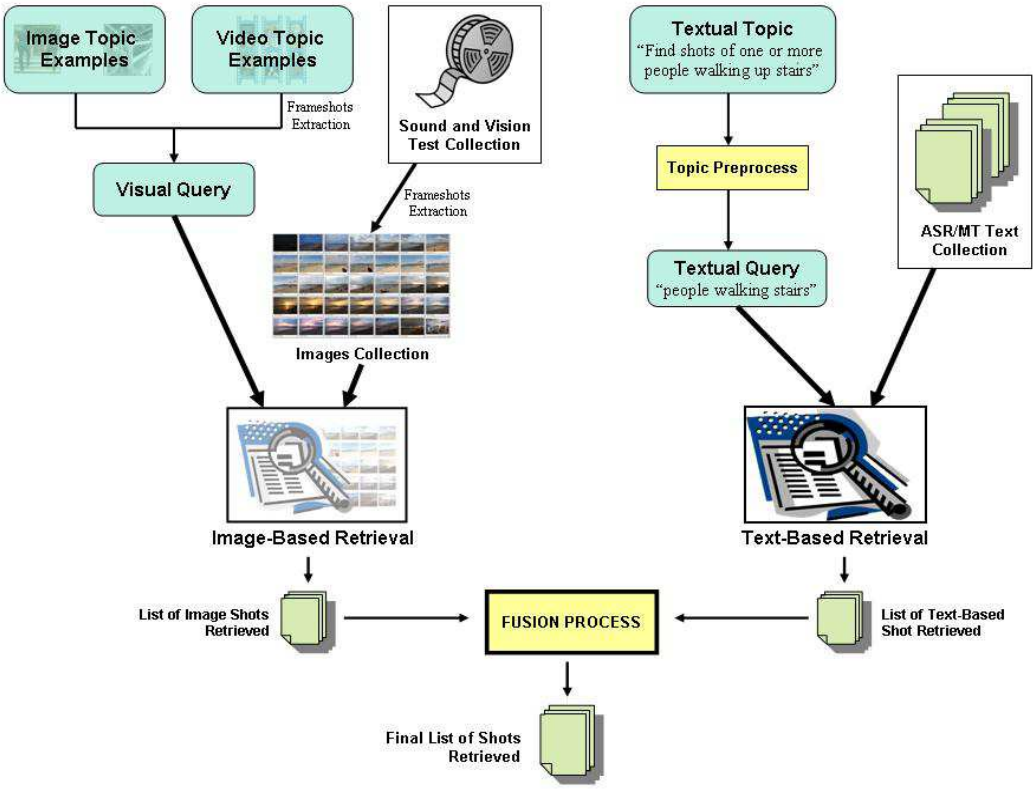

Figure 1: SINAI System Overview

2. To obtain the *audio segments* of each video file and to find what *video segments* belongs to each *audio segment*.

We have selected the first alternative for our system. In Figure 2 we can observe how the architecture takes all information about these *video segments* in every video file. Thus, each *video segment* includes all the *audio segments* that belong to it.

3.2 Image-based retrieval

For the image-based retrieval we have used GIFT as content based image retrieval tool. In order to use this tool, firstly we have to index all the images from the test collection.

For this subtask we have not applied any feature extraction strategy. Before executing the indexing of the image collection, we have extracted the *frameshots* of the test videos. Therefore, we are going to obtain a representative image from each segment of video (*mastershot*). We have calculated the midpoint of each *mastershot*, using start and end time, for extracting the image of the center of the scene. In order to generate the *videoshots* we have used the *ffmpeg*[5] program.

Making use of the *ffmpeg* tool, we have generated an image per second from video examples of the topics. All these generated images and the image examples of the topics, have been added to the test collection. Once we have prepared the test collection of images, the GIFT tool indexes them. This process is needed because GIFT does not allow to use images in queries that have not been indexed. The system filters the image examples and the images generated from video examples of the topics from the list of image shots retrieved by GIFT.

[5]FFMPEG multimedia system. Sourceforge. http://ffmpeg.sourceforge.net

```
<?xml version="1.0" encoding="UTF-8"?>
<Video id="TRECVID2007_105">
<VideoFile>BG_38907.mpg</VideoFile>
<Title>ZEMBLA</Title>
<MediaTime>
        <Duration hours="00" minutes="35" seconds="59" nf="22" nf_second="25"></Duration>
</MediaTime>
<Speakers total="26"></Speakers>
<VideoSegments total="143">
        <VideoSegment id="shot105_1">
                <ImageFile>BG_3890701.jpg</ImageFile>
                <MediaTime begintime="0.0" endtime="21.14">
                        <MediaTimePoint>T00:00:00:0F25</MediaTimePoint>
                        <Duration hours="00" minutes="00" seconds="21" nf="14" nf_second="25"></Duration>
                </MediaTime>
                <AudioSegments total="4">
                        <speech label="SPK01-001" begintime="2.71" endtime="6.57">
                                <real_time milliseconds="57761" frames="386" RTF="14.9640"/>
                                <text> chili is 't </text>
                        </speech>
                        [...]
                </AudioSegments>
        </VideoSegment>
        [...]
</VideoSegments>
</Video>
```

Figure 2: XML architecture example

The system provides visual queries for each topic. These queries were composed of the image examples and the extracted *frameshots* from video examples of the topics.

3.3 Fusion process

In the fusion process, the system mixes the list of image shots retrieved by GIFT and the list of text-based shot retrieved by LEMUR.

In order to generate the results of different experiments, we have proposed to not include the common *id shots* retrieved of both lists first. This is because the number of retrieved common *id shots* is low and it would make worse the final results.

Depending on the experiment, we have used different heuristics to generate the final list of retrieved shots in the fusion process. Basically, these heuristics are two:

- To include one shot from each sublist alternately, know as the '*RoundRobin*' strategy.
- To compute the final Ranking Status Value (RSV)[8] from the RSV of the item in both sublists by weighting them as follows:

$$RSV_i = \alpha \cdot RSV1_i + (1 - \alpha) \cdot RSV2_i$$

where RSV_1 is the i normalized score of shot retrieved by LEMUR, RSV_2 is the i normalized score of shot retrieved by GIFT and α is a constant.

4 Experiments and Results

The SINAI group has submitted a total of six runs for the automatic search task. As an overview, the mean average precision (MAP) values for our submitted runs are shown in Table 1. The run

Submission ID	Topic Expansion	Text Data	Visual Data	**MAP**
F_C_1_SINAI_1	no	yes	no	0.011
F_C_1_SINAI_2	yes	yes	no	0.010
F_C_1_SINAI_3	no	no	yes	0.000
F_C_1_SINAI_4	no	yes	yes	0.008
F_C_1_SINAI_5	no	yes	yes	0.008
F_C_1_SINAI_6	yes	yes	yes	0.008

Table 1: Overview of the runs in the automatic search task

F_C_1_SINAI_1 constitutes the required baseline run using only the query text and text features and the run **F_C_1_SINAI_3** constitutes the another required baseline run using only the visual features.

The query text has been first preprocessed and stemmed[11], and unnecessary words have been discarded by using a *stoplist* and removing the standard beginning of the query phrase. In the experiments we have only examined the query or term expansion used in the text topics. For each query text, we have expanded it with WordNet[6][12][9]. The topic expansion process locates the names and the verbs in the topic, and looks for three synonyms in WordNet for them. These synonyms are added to the text topic for its expansion. The run **F_C_1_SINAI_2** is the same that text baseline run but with term expansion.

For our text and visual runs, we have combined a selected output of the text-based retrieval subsystem with the output of the image-based retrieval subsystem. In the **F_C_1_SINAI_4**, we have mixed the list of text-based shots retrieved and the list of the image shots recovered following the *RoundRobin* algorithm. In the **F_C_1_SINAI_5** and **F_C_1_SINAI_6** experiments, we have mixed both lists using the RSV heuristic explained in previous section. The used value of α constant has been 75. In the **F_C_1_SINAI_5** the topics are without text expansion and in the **F_C_1_SINAI_6** the topics are with term expansion.

5 Conclusions and Future work

In this section, the results of experiments are discussed. Based on the MAP values, we can observe that using visual data does not seem to improve the overall results compared to the text-only baseline. The MAP value obtained in the required baseline run using only the visual features indicates that the GIFT tool does not work very well. The MAP value of that experiment is the worse result obtained.

On the other hand, the addition of synonyms to the text topics does not improve the text-only baseline result either.

The obtained MAP values of text and visual experiments do not equal the text-only baseline approach (0.011), the best score, because they include the 25 percent of best *shots* recovered by the image-based retrieval module.

The main conclusion about the performance of our runs is that we would have to use some visual feature extraction strategy to improve the overall results. On the analysis of images, also it would be interesting to use some concept ontology for multimedia data like LSCOM or some search algorithm based on concepts. On the text level, we could define a previous topic classification based on such concepts used for image-based retrieval.

6 Acknowledgments

This work has been supported by Spanish Government (MCYT) with grant TIN2006-15265-C06-03.

[6] http://wordnet.princeton.edu/

References

[1] M.C. Díaz-Galiano, M.A. García-Cumbreras, M.T. Martín-Valdivia, A. Montejo-Ráez, and L.A. Ureña-López. Sinai at imageclef 2006. In *In Proceedings of the Cross Language Evaluation Forum (CLEF 2006).*, 2006.

[2] M.C. Díaz-Galiano, M.A. García-Cumbreras, M.T. Martín-Valdivia, A. Montejo-Ráez, and L.A. Ureña-López. Sinai at imageclef 2007. In *In Proceedings of the Cross Language Evaluation Forum (CLEF 2007).*, 2007.

[3] M. Grubinger, P. Clough, A. Hanbury, and H. Müller. Overview of the imageclef 2007. photographic retrieval task. In *Working Notes of the 2007 CLEF Workshop. Sep, 2007. Budapest, Hungary.*

[4] Marijn Huijbregts, Roeland Ordelman, and Franciska de Jong. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of the second international conference on Semantics And digital Media Technologies (SAMT)*, Lecture Notes in Computer Science, Berlin, December 2007. Springer Verlag.

[5] M. Kaskziel and J. Zobel. Passage retrieval revisited. In *In Proceedings of the 20th annual International ACM Philadelphia SIGIR*, pages 178–185, 1997.

[6] M.T. Martín-Valdivia, M.A. García-Cumbreras, M.C. Díaz-Galiano, L.A. Ureña-López, and A. Montejo-Ráez. Sinai at imageclef 2005. In *In Proceedings of the Cross Language Evaluation Forum (CLEF 2005).*, 2005.

[7] M.T. Martín-Valdivia, F. Martínez-Santiago, and L.A. Ureña-López. Merging strategy for cross-lingual information retreival sistems based on learning vector quantization. In *Neural Processing Letters. Springer*, 2005.

[8] F. Martínez-Santiago, M.A. García-Cumbreras, M.C. Díaz-Galiano, and L.A. Ureña-López. Sinai at clef 2004: Usign machine translation resources with mixed 2-step rsv merging algorithm. In *In Proceedings of the Cross Language Evaluation Forum (CLEF 2004).*, 2004.

[9] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Five papers on wordnet. In *Special Issue of the International journal of lexicography*, 1993.

[10] H. Müller, T. Deselaers, E. Kim, J. Kalpathy-Cramer, T.M. Deserno, P. Clough, and W. Hersh. Overview of the imageclefmed 2007 medical retrieval and annotation tasks. In *Working Notes of the 2007 CLEF Workshop. Sep, 2007. Budapest, Hungary.*

[11] M.F. Porter. An algorithm for suffix stripping. In *Program 14*, pages 130–137, 1980.

[12] Jason Rennie. Wordnet::querydata: a Perl module for accessing the WordNet database. http://people.csail.mit.edu/ jrennie/WordNet, 2000.

[13] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.