Classification of Dual Language Audio-Visual Content: Introduction to the VideoCLEF 2008 Pilot Benchmark Evaluation Task

Martha Larson
ISLA
University of Amsterdam
The Netherlands
m.a.larson@uva.nl

Eamonn Newman
Centre for Digital Video Processing
Dublin City University, Ireland
enewman@computing.dcu.ie

Gareth J. F. Jones
Centre for Digital Video Processing
Dublin City University, Ireland
gjones@computing.dcu.ie

ABSTRACT

VideoCLEF is a new track for the CLEF 2008 campaign. This track aims to develop and evaluate tasks in analyzing multilingual video content. A pilot of a *Vid2RSS* task involving assigning thematic class labels to video kicks off the VideoCLEF track in 2008. Task participants deliver classification results in the form of a series of feeds, one for each thematic class. The data for the task are dual language television documentaries. Dutch is the dominant language and English-language content (mostly interviews) is embedded. Participants are provided with speech recognition transcripts of the data in both Dutch and English, and also with metadata generated by archivists. In addition to the classification task, participants can choose to participate in a translation task (translating the feed into a language of their choice) and a keyframe selection task (choosing a semantically appropriate keyframe for depiction of the videos in the feed).

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software: Performance evaluation

General Terms

Measurement, Performance, Experimentation, Languages

Keywords

Multimedia analysis, classification, speech recognition, benchmark test, cross-language information retrieval

1. INTRODUCTION

In 2008, a VideoCLEF track has been introduced at the Cross Language Evaluation Forum (CLEF)[1]. The VideoCLEF track for 2008 is being organized as a pilot of the *Vid2RSS*[2] task, which involves a number of subtasks. The results are delivered as RSS-feeds[3]. The main task is classifying videos into thematic classes (e.g., Architecture, Chemistry, History). The auxiliary tasks are translating feeds (i.e., the textual information contained in a feed) into another language and semantic keyframe selection. The track pursues the long term objective of developing the pipeline of processing necessary for fully automatic generation of topic-based feeds, specific to a particular information need and personalized to a particular language preference. The video data for Vid2RSS 2008 are Dutch television documentaries, which, while containing spoken Dutch content as the matrix language, also contain a high proportion of spoken English (i.e., interview guests often speak in English). A major focus of the Vid2RSS task is to optimize classification for videos whose spoken content is inherently multilingual.

CLEF pursues the goal of supporting research and development in information access to multilingual content. It promotes the development of retrieval infrastructures and also test suites of data for benchmarking that can be used and re-used by system developers. Previous and ongoing tasks have worked with multilingual news collections, scientific data, question answering, web data, speech data, geographic search, image search, and interactive retrieval and information processing [1]. The VideoCLEF task is designed to extend the Cross-Language Speech Retrieval (CL-SR) run at CLEF in recent years [3] to the broader challenge of search for video data.

VideoCLEF attempts to complement the TRECVid benchmark [4] by placing its main emphasis on exploiting spoken content (via speech recognition transcripts) and metadata associated with videos. Although participants are free to use features derived from the visual track of the video, it is not a required by the the task.

RSS-format was chosen so that system output can be displayed using a feed reader, which means that the runs performed by the partners are immediately visualizable and can be assessed by users, for example the archive staff. In this way, dissemination and transfer of task achievements is promoted. Grouping data into thematic categories is a task familiar to staff of large archives, namely to prepare a dossier of available resources on a certain topic for use by journalists and editors who are creating new content for broadcast.

Vid2RSS is motivated by the existence of video archives predominantly in one language, but containing embedded content in another. The extensive use of English-language interviews in Dutch-language programming means that Dutch media archives are a rich source of English language content for information seekers. Further, dual language programming offers a unique scientific opportunity. The two languages occur side by side in the program, both contributing to the spoken content of the program. Unlike other multilingual collections, the spoken content of the two languages is tightly intertwined and one does not duplicate the other. Dual language video thus presents the challenge of how to exploit speech features from both languages.

The balance of this paper is dedicated to a more detailed description of the Vid2RSS data collection and to explanation of the main task of Vid2RSS, the classification task.

[1] `http://www.clef-campaign.org`

[2] `http://ilps.science.uva.nl/Vid2RSS`

[3] Feeds in RSS, Real Simple Syndication, format

SSCS'08 SIGIR 2008 workshop, July 24, Singapore
.

2. VID2RSS 2008 TASK

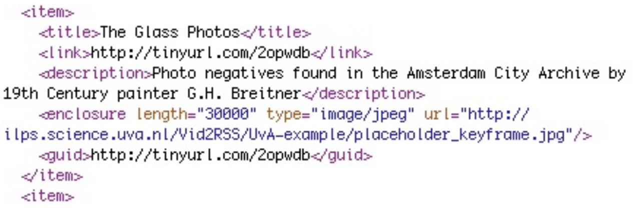
2.1 Task data

The Vid2RSS 2008 task data comprise 50 episodes (30 hours) of dual language television programs that are predominantly documentaries. The task data are broken down into a development set (10 episodes) and a test set (40 episodes). The video data have been supplied by Beeld & Geluid[4] one of the largest audio/video archives in Europe. Beeld & Geluid is located in Hilversum, Netherlands and is also referred to as Netherlands Institute of Sound and Vision. Dutch is the matrix language and English the embedded language (i.e., the documentaries are in Dutch and the experts interviewed in the documentaries speak English). The videos contain unplanned interview speech and at times challenging background conditions, such as music. Interview guests are not professional speakers and are often non-natives. Additional embedded languages, e.g., German and French, occur occasionally.

Automatic speech recognition (ASR) transcripts (first best hypothesis; MPEG-7 format) were kindly generated by the University of Twente using independent Dutch and English ASR-systems [2]. Each episode was transcribed in its entirety by each recognizer; no language detection was applied. Keyframes with location time stamps extracted from the video data stream were generated by Dublin City University [5] and are also made available to the participants, since keyframes are necessary to create the RSS-feeds, but keyframe selection is not a mandatory task. The archival metadata record is also supplied with each video in the collection. This record contains title and description fields as well as information about series, creator, publisher, rights and broadcast dates. The contents of the subject fields have been deleted, since these are the thematic class labels Vid2RSS aims to automatically generate.

2.2 Task output

The output format is a set of RSS-feeds, one for each topic class. The RSS-feed format is trivial to generate and is used as an output format because it can be easily visualized in a feed reader such as Netvibes[5]. The RSS-feeds are created by concatenating feed items, cf. Figure 1, for the video episodes assigned to a given topic class.



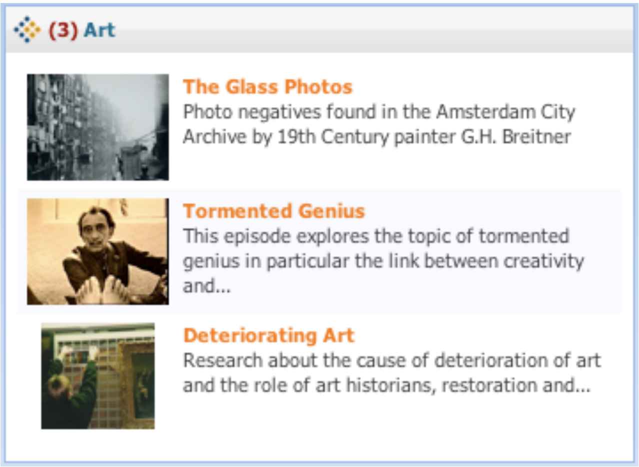
```
  <item>
    <title>The Glass Photos</title>
    <link>http://tinyurl.com/2opwdb</link>
    <description>Photo negatives found in the Amsterdam City Archive by
19th Century painter G.H. Breitner</description>
    <enclosure length="30000" type="image/jpeg" url="http://
ilps.science.uva.nl/Vid2RSS/UvA-example/placeholder_keyframe.jpg"/>
    <guid>http://tinyurl.com/2opwdb</guid>
  </item>
  <item>
```

Figure 1: Example of feed item, the building block of a feed

The feed items contain title and description fields taken from the archival metadata and are supplied with the task data.

2.3 The Classification Task

One step in the process of annotating a video for archival is to assign it one or more thematic class subject labels. Since archive staff use class labels for annotation and retrieval, the class labels automatically generated by the Vid2RSS classification task have concrete potential to support video search in an archive setting, and arguably also beyond. Ten thematic categories have been chosen for use in Vid2RSS 2008 (Archeology, Architecture, Chemistry, Dance, Film, History, Music, Paintings, Scientific research and Visual arts). These labels are a small subset of the class label inventory used by archive staff to annotate video for archival at Beeld & Geluid. We use archivist-assigned class labels contained in the subject fields of the metadata record as a gold standard for evaluating the Vid2RSS classification results. Participants are supplied with the topic labels and use them to collect their own training data.



(3) Art

The Glass Photos
Photo negatives found in the Amsterdam City Archive by 19th Century painter G.H. Breitner

Tormented Genius
This episode explores the topic of tormented genius in particular the link between creativity and...

Deteriorating Art
Research about the cause of deterioration of art and the role of art historians, restoration and...

Figure 2: Visualization of example topic feed in Netvibes

3. CONCLUSIONS AND OUTLOOK

The 2008 Vid2RSS task is intended as a pilot exercise to enable us to better understand the issues raised by the task of search in multilingual video content. Experience accrued in 2008 will inform the development of new tasks for the following years. In the future, would like to scale up the task, e.g., with use of a larger video corpus and an extended list of topic class labels. Also under consideration are the inclusion of embedded languages beyond English and the integration of additional sources of information such as subtitles, speaker change boundaries, speaker prosody, audio events and language identification. Further, introduction of new sub-tasks is envisaged, e.g., selection of a representative series of keyframes, feed generation on the basis of ad hoc queries and summarization aimed at automatic generation of description elements.

4. ACKNOWLEDGEMENTS

This research was supported by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

5. REFERENCES

[1] M. Braschler et al. From CLEF to TrebleCLEF: promoting Technology Transfer for Multilingual Information Retrieval. Second DELOS Conference on Digital Libraries, 2007.

[2] M. Huijbregts, R. Ordelman and F. de Jong Annotation of Heterogeneous Multimedia Content Using Automatic Speech Recognition. Proceedings of SAMT, 2007.

[3] P. Pecina, P. Hoffmannova, G. J. F. Jones, Y. Zhang and D. W. Oard, Overview of the CLEF 2007 Cross-Language Speech Retrieval Track, Proceedings of the CLEF 2007 Workshop, 2007.

[4] A. F. Smeaton, P. Over and W. Kraaij, Evaluation Campaigns and TRECVid, In Proceedings of MIR 2006 - 8th ACM SIGMM International Workshop on Multimedia Information Retrieval, 2006.

[5] J. Calic, S. Sav, E. Izquierdo, S. Marlow, N. Murphy and N. O'Connor Temporal Video Segmentation for Real-Time Key Frame Extraction Proceedings of ICASSP 2002.

[4] http://www.beeldengeluid.nl

[5] http://www.netvibes.com