Overview of VideoCLEF 2008: Automatic generation of topic-based feeds for dual language audio-visual content

Larson, M.; Newman, E.; Jones, G.

Publication date
2008

Published in
Working Notes for the CLEF 2008 Workshop: 17-19 September, Aarhus, Denmark

Overview of VideoCLEF 2008: Automatic Generation of Topic-based Feeds for Dual Language Audio-Visual Content

Martha Larson[1] Eamonn Newman[2] Gareth Jones[2]

[1]ISLA, University of Amsterdam
[2]CDVP School of Computing, Dublin City University
m.a.larson@uva.nl, {eamonn.newman | gareth.jones}@computing.dcu.ie

Abstract

The VideoCLEF track, introduced in 2008, aims to develop and evaluate tasks related to analysis of and access to multilingual multimedia content. In its first year, VideoCLEF piloted the Vid2RSS task, whose main subtask was the classification of dual language video (Dutch-language television content featuring English-speaking experts and studio guests). The task offered two additional discretionary subtasks: feed translation and automatic keyframe extraction. Task participants were supplied with Dutch archival metadata, Dutch speech transcripts, English speech transcripts and 10 thematic category labels, which they were required to assign to the test set videos. The videos were grouped by class label into topic-based RSS-feeds, displaying title, description and keyframe for each video.

Five groups participated in the 2008 VideoCLEF track. Participants were required to collect their own training data; both Wikipedia and general web content were used. Groups deployed various classifiers (SVM, Naive Bayes and k-NN) or treated the problem as an information retrieval task. Both the Dutch speech transcripts and the archival metadata performed well as sources of indexing features, but no group succeeded in exploiting combinations of feature sources to significantly enhance performance. A small scale fluency/adequacy evaluation of the translation task output revealed the translation to be of sufficient quality to make it valuable to a non-Dutch speaking English speaker. For keyframe extraction, the strategy chosen was to select the keyframe from the shot with the most representative speech transcript content. The automatically selected shots were shown, with a small user study, to be competitive with manually selected shots. Future years of VideoCLEF will aim to expand the corpus and the class label list, as well as to extend the track to additional tasks.

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; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Classification, Translation, Keyframe Extraction, Speech Recognition, Evaluation, Benchmark, Video

1 Introduction

VideoCLEF was a new track piloted at CLEF 2008.[1] The goal of the track is to develop and evaluate tasks involving the analysis of multilingual video content. In particular, we are interested in *dual language* video. Dual language video is video content in which two languages are spoken, but the content of one does not duplicate (i.e., is not a translation of) the content of the other. Prime examples of dual language video content are documentaries and talk shows where interviewees and studio guests do not speak the dominant language of the show (referred to as the *matrix language*), but rather speak another language (referred to as the *embedded language*). The VideoCLEF task was introduced as the successor to the Cross-Language Speech Retrieval (CL-SR) run at CLEF from 2005 to 2007 [6]. The goal is to extend the achievements of CL-SR to the broader challenge of search for video data. VideoCLEF is intended to complement the TRECVid benchmark [7] by emphasizing the exploitation of spoken content (via speech recognition transcripts) and also of archival metadata associated with videos. While TRECVid concerns itself with what is depicted in a video, VideoCLEF focuses on what is described in a video, in other words, what a video is about. VideoCLEF participants are free to use features derived from the visual track of the video, but it is not a required aspect of the task.

1.1 Data

The video data for VideoCLEF 2008 was supplied by the Netherlands Institute of Sound and Vision[2] (called in Dutch *Beeld & Geluid*), one of the largest audio/video archives in Europe. The dual language content contained in the Sound and Vision archives provided the initial inspiration for the VideoCLEF 2008 task. Although the dominant spoken language of much Dutch television programming is, not surprisingly, Dutch, many other languages are spoken. Dutch television is subtitled rather than dubbed. The extensive use of English and other languages in interviews and studio discussions in Dutch television programming means that Dutch media archives are a rich source of spoken content in languages other than Dutch.

Dual language content is an interesting subject of research investigation for two reasons. First, as mentioned above, in dual language content, two or more languages exist side by side. The languages are intertwined, but not duplicated. Each spoken language represents a separate source of evidence for semantic analysis, classification and retrieval of video. Although we limited VideoCLEF 2008 to two languages, the natural extension of the task is to involve as information sources all languages present in the video content. In the Sound and Vision archive, additional languages include not only other European languages, but also a mixture of languages from the other continents. Further, dual language video also implies the presence of subtitles, which (again, this was not yet done in the pilot year 2008) are a valuable further source of semantic evidence. Second, dual language content is useful to information seekers who do not speak the dominant language of the archive. Dutch documentaries are of high quality and media archives contain valuable information nuggets in the form of interviews with historically significant figures. VideoCLEF 2008 aimed to take a first step towards providing access to non-Dutch content hidden within a predominantly Dutch language video collection.

1.2 Tasks

In 2008, VideoCLEF consisted of one task, called *Vid2RSS*.[3] A supplementary description of the Vid2RSS task can be found in [4]. The main subtask of this task was a classification task involving automatically assigning thematic subject category labels to dual language video. The classification task was chosen since it is a straightforward classic video analysis task with a high potential for application in real world systems. Thematic subject labels can be understood to be high-level semantic features. Such features can be applied directly in a faceted browsing system or they can be used to support retrieval or other video analysis tasks downstream. The subject labels used for Vid2RSS have known utility for multimedia search. They are a subset of classes used by archive staff for archival and retrieval at Sound and Vision. The creation of groups of resources related to one topic is a familiar task to the staff of large archives, who are often called

[1] http://www.clef-campaign.org
[2] http://www.beeldengeluid.nl
[3] http://ilps.science.uva.nl/Vid2RSS

upon to create a dossier on a particular topic for use in production of new content for broadcast. The choice of the classification task as a task for VideoCLEF was also influenced by an important practical consideration; archivist assigned subject labels are available for the test data and provide the gold standard for task performance evaluation.

Participants submitted the Vid2RSS results as a series of topic-based RSS-feeds. The feeds are trivial to generate. Generation involves concatenating feed item elements corresponding to the videos that have been assigned a certain class label. The feed item elements contain the title of the video, a short description and a representative keyframe and were supplied with the test data. The purpose of requiring output in RSS-feed format was to make the results of the runs submitted by the different sites easily visualizable. RSS-feeds can be displayed in a feedreader and can be easily assessed by end users, for example archive staff. By using RSS as the output format, we hope that we can narrow the distance that must be traversed between experimental runs in a benchmark campaign and exploitation of results achieved in a real-world application.

In addition to the classification subtask, which was mandatory, participants could also carry out two additional subtasks, a translation subtask and a keyframe extraction subtask. The following sections of the paper describe each of the tasks in turn, summarizing the approaches chosen by the individual participants and the task results. The paper finishes with a conclusion and outlook.

2 Classification Task

The goal of this task was to reproduce the subject labels that were hand assigned to the test set videos by archivists at Sound and Vision. Ten thematic categories were chosen, representing a small subset of the subject labels in use at Sound and Vision: Archeology (archeologie), Architecture (architectuur), Chemistry (chemie), Dance (dansen), Film (film), History (geschiedenis), Music (muziek), Paintings (schilderijen), Scientific research (wetenschappelijk onderzoek) and Visual arts (beeldende kunst).

For each video, the task participants were provided with archival metadata including the description and title of the video. As mentioned above, subject labels were removed from this archival metadata record. Participants received speech transcripts from both languages. The speech transcripts included the first best hypothesis of the recognizer and were encoded in MPEG-7 format. The transcripts were generated by the University of Twente [2]. No language detection was used, so both the Dutch and the English transcript reflect a recognition of the video in its entirety. The required task was to perform classification making use of the speech recognition transcripts only.

2.1 Techniques

2.1.1 Chemnitz University of Technology (CUT)

The Chemnitz University of Technology (CUT) team chose to carry out the task using a Naive-Bayes and a k-nearest neighbor (k-NN) classifier. They derived training data for the classifiers from identically or similarly named categories in the English and the Dutch Wikipedia. In their experiments, they varied the composition of the feature set (i.e., the vocabulary of terms) used for classification. Stemming and stopword removal were applied. The results suggest that it is helpful to eliminate terms that occur in multiple classes. Also, depth to which they descended into the Wikipedia category while gathering data impacted results. The performance achieved by their method on the development data, unfortunately did not transfer to the test data. In particular, the CUT team notes that classification performance did not improve when the archival metadata was added to the mix. Performance on the combination of archival metadata and transcripts remained comparable to performance on transcripts alone.

2.1.2 Dublin City University (DCU)

The Dublin City University (DCU) team approached the task as an information retrieval problem and used an off-the-shelf Information Retrieval system implementing the vector space model. Both stopword removal and stemming were applied in the feature extraction step. The label of each subject category was used as a query. The DCU team experimented with two dimensions: (1) limiting the recall of the task by

labeling a video only with the most specific category label that retrieved it, and, (2) using blind relevance feedback to expand the label of the subject category into a richer query. The Dutch speech transcripts used alone were more useful than the English speech transcripts used alone. Using metadata alone allowed the system to achieve high precision, but did not out-perform the run using Dutch speech transcripts alone.

2.1.3 MIRACLE Research Consortium (MIRACLE)

The MIRACLE Research Consortium (MIRACLE) chose a classifier based on the k-nearest neighbor algorithm. Representations of the video episodes were used as queries to perform retrieval on a knowledge base containing Wikipedia articles. Each episode was assigned the label that was associated with the most retrieved Wikipedia articles. In the experiments, the length of the results list was set to 10. Stopword removal and stemming were applied. The MIRACLE team hypothesized that performance is improved in cases where there are a larger number of Wikipedia articles of the appropriate class available in the knowledge base.

2.1.4 University of Amsterdam (UAms)

The University of Amsterdam (UAms) team picked a Support Vector Machine with a linear kernel to use as the classifier. They applied χ^2 feature selection; no stopword removal or stemming was performed. In order to collect training data, the class labels were submitted as a query to Dutch and English Wikipedia and the returned articles were used as the training set. Experimentation was performed with adding archival metadata to speech transcripts for the representation of test documents (improve performance) and combining Dutch and English speech transcripts (did not outperform use of Dutch speech transcripts by themselves).

2.1.5 University of Jaén (SINAI)

The SINAI team from the University of Jaén collected topical data from the internet by submitting the thematic class labels as queries to Google and harvesting the top 10 documents returned, which were amalgamated into a single document. One such document from each class was indexed. Stopword removal and stemming were applied. Retrieval was performed on this collection using the language modeling framework. The queries were derived from the speech transcripts and from the archival metadata. A video was assigned the label corresponding to the top ranked document.

2.2 Results

This section reports the results achieved on the classification task by all participating sites and comments on the techniques used and the trends observed. Results are reported in terms of micro-averaged f-scores and macro-averaged f-scores [3]. The f-score is the harmonic mean between precision and recall. The micro-average reflects a document-centric system performance and is calculated directly with respect to the entire collection. The macro-average reflects class-centric performance and is calculated by first computing the f-score for each individual topic class and then averaging over all classes.

The results of all runs are presented in Table 1. The top micro-averaged f-score was 0.53, achieved by SINAI with run SINAI-JEAN-Class-II and the top macro-averaged f-score was 0.59, achieved by DCU with run dcu_run4. It should be noted that good micro-averages and macro-averages might not reflect the type of performance that humans intuitively feel is best. Run dcu_run4 has a high macro-average since it sacrifices precision for recall in 6 of the 10 classes and sacrifices recall for perfect precision in the other 4. Run SINAI-JEAN-Class-II has a high micro-precision, due to the fact that it assigns class labels in only 3 of the ten classes and in these three it performs well. In both cases the performance scores are high, but it can be argued that such a classification strategy might not appeal to a human user who would like to have a chance to find videos in all 10 topic categories.

Humans might actually find the runs dcu_run1 and MIRACLE-CNLMeta to yield more usable performance. Here, the macro-precision and the micro-precision are better balanced. Note that these are two categories in which the improvement in system performance over multiple (although not all) competitor runs can be shown to be statistically significant at the $p <= 0.05$ level according to the Wilcoxon signed rank test.

RunID	micro-averaged f-score	macro-averaged f-score	feature language	test doc rep	site
CUT-C1R1▲	0.15	0.27	en/nl	asr	CUT
CUT-C1R2	0.11	0.14	en/nl	asr	CUT
CUT-C2R1	0.13	0.26	en/nl	asr/md	CUT
CUT-C2R2	0.13	0.17	en/nl	asr/md	CUT
dcu_run1▲	0.41	0.54	nl	asr	DCU
dcu_run2▲	0.25	0.47	en	asr	DCU
dcu_run3▲	0.28	0.58	nl	asr	DCU
dcu_run4	0.28	**0.59**	en	asr	DCU
dcu_run5	0.29	0.43	nl	md	DCU
MIRACLE-CNL	0.46	0.49	nl	asr	MIRACLE
MIRACLE-CNLEN	0.39	0.27	nl/en	asr	MIRACLE
MIRACLE-CNLMeta▲	0.47	0.47	nl	asr/md	MIRACLE
uams08m	0.18	0.17	nl	md	UAms
uams08asrd	0.10	0.41	nl	asr	UAms
uams08masrd	0.15	0.45	nl	asr/md	UAms
uams08asrde	0.09	0.14	nl/en	asr	UAms
uams08masrde	0.09	0.33	nl/en	asr/md	UAms
SINAI-Class-I	0.51	0.49	nl	asr	SINAI
SINAI-Class-II	**0.53**	0.51	en	asr	SINAI
SINAI-Class-I-Trans	0.10	0.40	nl	md	SINAI

Table 1: Evaluation results for all runs from all participants (nl = Dutch; en = English, asr = Automatic Speech Recognition transcripts; md= archival metadata; test. doc. rep. = source of the features for the test document representation). Runs with a statistically significant improvement over at least one other competitor run are indicated by ▲ (Wilcoxon signed rank test; $p <= 0.05$).

2.2.1 Runs using speech recognition transcripts only

Runs that used speech transcripts alone were competitive with runs that used archival metadata alone or combined archival metadata with speech transcripts. These results indicate that there is potential for automatically assigning thematic category labels for videos that lack archival metadata.

2.2.2 Runs integrating English speech transcripts with Dutch information sources

No participant was able to exploit the speech recognition transcripts for the English language in order to improve performance. Participants conjectured that this might have been due to the fact that there was more Dutch spoken content in the documentaries than there was English spoken content or that the English speech recognition transcripts had a higher word error rate than the Dutch speech recognition transcripts. In the future, we would like to try segmenting the videos using a language detector so that transcripts for English are generated only where English is spoken in the video.

2.2.3 Top performing classes

The breakdown of the performance of the runs over the individual thematic categories is shown in Table 2. It can be seen that *Music* is the class for which the best performance was achieved, cf., SINAI-Class-II, MIRACLE-CNLEN, MIRACLE-CNLMeta. It should be noted that this is also the class with the highest number of videos in the test corpus. The fact that relatively high performance levels could be attained for individual classes suggests that progress can still be achieved on the classification task if more research and development effort is devoted to it in the future.

2.2.4 Comments on the evaluation metric

We would like to mention here why the metrics chosen for VideoCLEF 2008 may not be adequate to reflect all relevant aspects of system performance. A high micro-average does not reflect how system performance is distributed over classes. However, the macro-average also has its shortcomings. In order to calculate the macro-average, it is necessary to define its behavior in cases where, for a given class, there are no videos in the collection that belong to that class. This case leads to division by zero when calculating the recall. We defined recall for these classes to be 1.0, since there exists no video of this class in the collection to which the classifier has failed to assign the correct class label. It is also necessary to define behavior in cases where a system fails to assign any videos to a particular class. This case leads to division by zero when calculating the precision. We defined precision for these classes to be 1.0, since there are no videos in the collection to which the classifier has erroneously assigned this class label. An alternative solution would be to average precision and recall only over classes that don't give rise to a division by zero problem. This solution might encourage systems to artificially inflate precision by ignore difficult classes. We did not adopt this solution. An advantageous aspect of the Vid2RSS task is that the visualization of the results as RSS-feeds makes it easy for humans to grasp differences in run performance and to understand the mismatch between runs that might be most useful for real world applications and runs that achieve high performance. The visualization serves to compensate for evaluation metrics which do not present well-rounded picture of system performance.

RunID	**Arche**	**Archi**	**Chem**	**Dance**	**Film**	**Hist**	**Mus**	**Paint**	**Sci**	**Arts**
Raw count correct videos	**7**	**0**	**0**	**3**	**3**	**10**	**22**	**3**	**4**	**5**
CUT-C1R1▲	0.44	0.00	0.00	0.00	0.20	0.14	0.16	0.20	0.00	0.14
CUT-C1R2	0.15	0.00	0.00	0.00	0.00	0.10	0.21	0.17	0.00	0.15
CUT-C2R1	0.44	0.00	0.00	0.00	0.20	0.14	0.08	0.20	0.00	0.13
CUT-C2R2	0.15	0.00	0.00	0.18	0.00	0.10	0.21	0.22	0.00	0.17
dcu_run1▲	0.60	1.00	0.00	0.50	0.46	0.24	0.47	0.57	0.40	0.00
dcu_run2▲	0.25	0.00	1.00	0.18	0.15	0.50	0.17	0.40	0.00	0.00
dcu_run3▲	0.30	0.00	1.00	1.00	0.14	0.40	0.71	0.14	0.18	0.00
dcu_run4	0.00	0.00	1.00	0.14	0.14	0.40	0.71	0.14	0.00	0.00
dcu_run5	0.00	1.00	1.00	0.00	0.33	0.18	0.53	0.00	0.00	0.00
MIRACLE-CNL	0.18	1.00	0.00	0.00	0.00	0.27	0.76	0.00	0.44	0.00
MIRACLE-CNLEN	0.18	0.00	0.00	0.00	0.29	0.34	0.79	0.00	0.35	0.27
MIRACLE-CNLMeta▲	0.33	0.00	0.00	0.22	0.00	0.46	0.79	0.00	0.35	0.17
uams08m	0.00	0.00	0.00	0.00	0.09	0.38	0.44	0.00	0.15	0.00
uams08asrd	0.25	0.00	0.00	0.00	0.06	0.18	0.00	0.00	0.33	0.00
uams08masrd	0.26	1.00	1.00	0.11	0.00	0.00	0.00	0.00	0.18	0.14
uams08asrde	0.17	0.00	0.00	0.00	0.00	0.27	0.00	0.11	0.11	0.22
uams08masrde	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.00
SINAI-Class-I	0.00	1.00	1.00	0.00	0.00	0.18	0.79	0.00	0.33	0.00
SINAI-Class-II	0.00	1.00	1.00	0.00	0.00	0.17	0.81	0.00	0.57	0.00
SINAI-Class-I-Trans	0.00	1.00	1.00	0.00	0.00	0.00	0.16	0.00	0.16	0.00

Table 2: F-scores of each run reported for each individual class. Runs with a statistically significant improvement over at least one other competitor run are indicated by ▲ (Wilcoxon signed rank test; $p <= 0.05$). Full names of the thematic categories are: Archeology, Architecture, Chemistry, Dance, Film, History, Music, Paintings, Scientific research and Visual arts.

3 Translation Task

The translation task, which was a discretionary task, required participants to translate topic-based feeds from Dutch into a target language. The feeds consist of concatenations of feed items, each describing

a video with that video's title, a small description derived from the archival metadata and a keyframe representing the video's content. One participant, CUT, carried out the translation task. CUT chose to translate the feeds into English and to use Google's AJAX language API.

Evaluation of the feeds was carried out using human assessment of adequacy and fluency performed by 3 assessors. All assessors had high-level mastery of both the source and target language. The assessment procedure was adapted from the TIDES *Specifications for human assessment of translation quality*.[4] Assessors were asked to assess fluency and adequacy of the translation of the feed item metadata (title and description) for each video a five point scale. For fluency, they were asked to answer the question *How do you judge the fluency of this translation?* and assign points on the basis of the following answers: 5 = Flawless English, 4 = Good English, 3 = Non-native English, 2 = Disfluent English, 1 = Incomprehensible. For adequacy, they were asked to answer the questions *How much of the meaning expressed in the original Dutch version of the video title and description is also expressed in the English translation?* and assign points on the basis of the following answers: 5 = All, 4 = Most, 3 = Much, 2 = Little, 1 = None. On average, assessors gave feed items a score of 2.82 for Fluency and 3.49 for adequacy.

One of the main problems with the translation is that compound words often failed to be translated. For example the Dutch word "tiendelige," which is a compound that means consisting of ten parts, is written simply as "tiendelige" in the English translation. The word "concertpianist" meaning concert pianist, is translated as only "concert" with mention of pianist dropped. Another problem is that proper names were translated in cases when then are homonyms with other words. Despite these glitches, on the whole the translation was satisfactory and certainly demonstrated potential to allowing non-Dutch speakers to understand the contents of the topic-based feeds.

4 Keyframe Extraction Task

Participants were provided with a segmentation of the videos into shots and a set of keyframes, one keyframe per shot. The segmentation and the shot level keyframe data was provided by Dublin City University [1]. The Vid2RSS keyframe extraction task required the participant to pick the keyframe from the provided set that best represented the semantic content of the video. Note that the task of automatically extracting a keyframe to represent a shot was not evaluated in VideoCLEF 2008. The set of keyframe level shots was taken as a given, and participants were required to chose the most appropriate keyframe from this set.

4.1 Keyframe Extraction Experiments and Results

Only one participant, MIRACLE, participated in the keyframe extraction task, which was discretionary. MIRACLE chose the keyframes based on the content of the speech recognition transcript associated with the shot. The MIRACLE team based their approach on the assumption that a representative shot for the video is a shot for which the spoken content is the least different from the spoken content of the video as a whole. They selected the keyframe of the shot whose speech recognition transcript vector has the closest cosine distance to the speech recognition transcript vector of the video as a whole.

Keyframe extraction was tested in a small scale user study in which subjects were given the title and description of a video and asked to chose between two candidates for a keyframe to represent the semantic content of that video. One candidate was the baseline human selected keyframe and the other was the keyframe automatically selected by MIRACLE. On an average 44% of the videos had automatic keyframes that were well selected, meaning that they were either identical to the manually selected keyframes (2 cases), or that the subjects preferred the automatically selected keyframe to the manual one. These results suggest that the automatic keyframe extraction is a very viable competitor with manual keyframe selection.

4.2 Keyframe Extraction Evaluation

During the course of the user study, several important trends emerged that should be mentioned here since they serve to illustrate how challenging the keyframe extraction task actually is and the limited ability of

[4] `http://projects.ldc.upenn.edu/TIDES/Translation/TransAssess04.pdf`

the evaluation score to reflect the level of challenge. First, subjects often have a very mild preference for one keyframe over the other, implying that when a subject chooses the manually selected keyframe over the automatically extracted keyframe, it does not mean that the automatically extracted keyframe was inappropriate. Second, subjects' preference of keyframe was dependent on their knowledge of the topic of the video. In the case of a documentary about Frank Zappa, subjects who could recognize Frank Zappa by sight chose the keyframe picturing him, and rejected the other keyframe, which was technically a much clearer picture. The same phenomenon was observed for a video about World War II. Subjects that could identify Churchill by sight preferred the keyframe picturing him. Third, the comments of the experimental subjects reflected that their picks were dependent on whether they felt that the keyframe should depict the genre (documentary) or particular television series, or whether they felt it should depict the novel content of the particular video episode. In some cases, familiarity with the television series to which the video belonged impacted the subject's decision on which keyframe to pick. Subjects commented on some occasions that they simply preferred the "prettier" keyframe. One subject preferred the keyframe that made the video seem more enticing. Finally, there were a lot of detail that subjects paid attention to. For example, in one case one keyboard shot was preferred above another because it was slightly shifted revealing knobs that showed the keyboard to be an electric one. Taking such detail into account will probably remain a challenge for semantic keyframe extraction until far into the future.

5 Conclusions and Future Plans

The Vid2RSS task in the VideoCLEF 2008 track involved classification, translation and keyframe extraction performed on dual language video. All in all, evaluation of the classification runs demonstrated that there is quite a bit of improvement left to be achieved on this task. However, strong performance by classifiers for particular thematic categories, especially classifiers for videos treating the topic of *Music*, leads us to believe that improved performance can be achieved in the future. Further, the classification task demonstrated that both Wikipedia and the Web at large to be promising sources of training data. Finally, simple approaches that recast the classification problem as an information retrieval task, yielded strong results.

The results of the discretionary translation task were satisfactory, although they would have been more revealing had more than a single site participated. A further comment should be made at this juncture concerning translation in the Vid2RSS task. Recall that the Vid2RSS task was originally motivated by the idea the multimedia archives contain multilingual content that is of high informational value if it can be made available to users who do not master the dominant language of the archive. The results of the translation task strongly suggest that the impasse for providing non-Dutch speakers access to usable content in a predominantly Dutch archive does not lie in the problem of providing usable translations of video titles and descriptions.

The results of the discretionary keyframe selection task were very encouraging. Here again, more elaborate conclusions would be supported had more than a single site participated. However, the small scale user study did demonstrate that automatically selected keyframes are competitive with manually selected keyframes. This result confirms the usefulness of the speech transcript associated with the video as a source of features for selecting a keyframe capable of semantically representing the video.

We were pleased with the success of the idea of having participants deliver their results as topic-based RSS-feeds. Feeds for the same class from different runs can be compared graphically in a feed reader with very minimal effort. Such a visualization makes it easier to get feedback on the usability of task results from potential end users who can gain a quick impression of the potential utility of the classification. The visualization aspect proved to be particularly important since, as mentioned above, we were not particularly convinced that the evaluation metrics chosen for the year's task truly reflected the potential usefulness of the results in a application.

We consider the VideoCLEF pilot track to have successfully demonstrated that the classification of dual language television documentaries into subject classes is a challenging and interesting task. In particular, we would like to note that the experiences of the pilot year of Vid2RSS strongly suggest that classification of video content is not always as easy as classification of broadcast news content, for which reasonable performance can be achieved in a relatively straightforward fashion [5]. We believe that a significant source of challenge lies in the fact that the videos contain a high proportion of unscripted speech in the form of

interviews and discussions. Associated with such speech, which can be characterized as conversational, is a wide vocabulary, potentially sparse in on-topic words, and an informal style including disfluencies and sentence fragments. Combination of features derived from multiple sources (speech transcripts of both matrix and embedded language and metadata, where available) seems to offer a line of investigation with solid potential to improve classification performance, although such improvement was not realized in the initial year of the VideoCLEF track.

In the future, we would like to continue the Vid2RSS task, expanding it to include additional classes and more test data. We would also like to be able to provide participants with training data, in order to understand the potential of better matched training data for improving classification performance. We would like to extend VideoCLEF to include a *quote retrieval task*. This task is modeled on the use case where an editor or journalist is searching for a clip in which a prominent personage says a certain famous phrase for purposes of reuse in a new production. Additionally, we would like to introduce a novel, exploratory task that will concentrate on classifying video according to categories which are more challenging than thematic categories since they do not bear a direct relationship to the video's semantic content. In particular, we envision a *favorites filter* task. The first formulation of the favorites filter task would be very simple: participants would be required to select among three videos the video that human assessors chose the most boring, the most superficial or the most outdated.

6 Acknowledgements

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

References

[1] J. Calic, S. Sav, E. Izquierdo, S. Marlow, N. Murphy, and N. O'Connor. Temporal video segmentation for real-time key frame extraction. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing ICASSP 2002, Orlando, Florida*, 2002.

[2] M. Huijbregts, R. Ordelman, and F. de Jong. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of SAMT*, 2007.

[3] P. Jackson. *Natural Language Processing for Online Applications*. Natural Language Processing. John Benjamins, Philadelphia, 2002.

[4] M. Larson, E. Newman, and G. Jones. Classification of dual language audio-visual content: Introduction to the VideoCLEF 2008 pilot benchmark evaluation task. In *Proceedings of the SIGIR 2008 Workshop on Searching Spontaneous Conversational Speech*, pages 71–72, 2008.

[5] G. Paass, E. Leopold, M. Larson, J. Kindermann, and S. Eickeler. SVM classification using sequences of phonemes and syllables. In *PKDD '02: Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 373–384, 2002.

[6] P. Pecina, P. Hoffmannova, G. J. F. Jones, Y. Zhang, and D. W. Oard. Overview of the CLEF 2007 cross-language speech retrieval. In *Proceedings of the CLEF 2007 Workshop*, 2007.

[7] A. F. Smeaton, P. Over, and W. 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.