LinkedTV at MediaEval 2013 Search and Hyperlinking Task

M. Sahuguet[1], B. Huet[1], B. Červenková[2], E. Apostolidis[3], V. Mezaris[3], D. Stein[4], S. Eickeler[4], J.L. Redondo Garcia[1], R. Troncy[1], and L. Pikora[2]

[1]Eurecom, Sophia Antipolis, France. [sahuguet,huet,redondo,troncy]@eurecom.fr

[2]University of Economics, Prague, Czech Republic. [barbora.cervenkova,lukas.pikora]@vse.cz

[3]Information Technologies Institute, Thessaloniki, Greece. [apostolid,bmezaris]@iti.gr

[4]Fraunhofer IAIS, Sankt Augustin, Germany. [daniel.stein,stefan.eickeler]@iais.fraunhofer.de

ABSTRACT

This paper aims at presenting the results of LinkedTV's first participation to the Search and Hyperlinking task at MediaEval challenge 2013. We used textual information, transcripts, subtitles and metadata, and we tested their combination with automatically detected visual concepts. Hence, we submitted various runs to compare diverse approaches and see the improvement when adding visual information.

1. INTRODUCTION

This paper describes the framework used by the LinkedTV team to tackle the problem of Search and Hyperlinking inside a video collection [2]. The applied techniques originate from the LinkedTV project[1], which aims at integrating TV and internet experience, by enabling the user to access additional information and media resources aggregated from diverse sources, thanks to automatic media annotation.

2. PRE-PROCESSING STEP

Concept detection was performed on the key-frames of the video, following the approach in [6], while the algorithm for Optical Character Recognition (OCR) described in [8] was used for text localization. Moreover, for each video, we extracted keywords from the provided subtitles, based on the algorithm presented in [9]. Finally, we grouped the predefined video shots into bigger segments (scenes), based on the visual similarity and the temporal consistency among them, using the method introduced in [7].

3. OUR FRAMEWORK

3.1 Lucene indexing

We indexed all available data in a Lucene index at different granularities: video level, scene level, shot level and segments created using sliding window algorithm [3]. Documents were represented by both textual fields (for a text search) and floating point fields (for the visual concepts).

3.2 From visual cues to detected concepts

Text search is straightforward with Lucene, by using the default text search based on TF-IDF values. In order to incorporate visual information to the search, we mapped keywords extracted from the visual cues query (using Alchemy API[2]) to visual concepts using a semantic word distance based on Wordnet synsets [5]. When visual concepts were detected in the query, we enriched the textual query by range queries on the values of the corresponding visual concepts.

3.3 Search task

We concatenated textual and visual queries to perform the text query. Two strategies were adopted: we either used segments indexed in the Lucene engine, or performed queries creating segments on the fly, by merging video segments based on their score.

Performing a text query on the video index often returned the relevant video in the top of the list. Hence, some runs first restrict the pool of videos that are going to be searched to a small number, and then perform additional queries for smaller segments inside this pool.

We submitted 9 runs in total:

- *scenes-C*: Scene search using textual and visual cues.
- *scenes-noC*: Same as previous using textual cues only (no visual cues) for comparison purposes.
- *part-sc-C*: Partial scenes search from shot boundary using textual and visual cues following three steps: filtering of the list of videos; querying for shots inside each video; ordering them by score. As a shot is a unit that is too small to be returned to a viewer, we completed the segment with the end of the scene that includes this shot.
- *part-sc-noC*: Same as previous using textual cues only.
- *cl10-C*: Temporal clustering of shots within a video using text and visual cues in the following manner: filtering out the set of videos to search; computing scores for every shot in the video; clustering together shots closer than 10 seconds apart (scores were added to form the final score).
- *cl10-noC*:Same as previous using text search only.
- *scenes-S or scenes-U or scenes-I*: Scene search using only textual cue from transcript or subtitle, no metadata.
- *SW-60-I or SW-60-S*: Search over segments created by the sliding window algorithm for LIMSI/Vocapia [4] tran-

[1] http://www.linkedtv.eu/

[2] http://www.alchemyapi.com/

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

scripts and subtitles, where the size of the sliding windows is 60.

- *SW-40-U*: Same as above for LIUM transcript with sliding window size of 40.

3.4 Hyperlinking task

A first approach consisted in reusing the search component with the scene approach and the shot clustering approach. A query was crafted from the anchor: the text query was made by extracting keywords from the subtitle aligned at start time and end time of the anchor. Visual concepts scores were extracted from the keyframes of shots contained in the anchor. If the anchor was constituted by more than one shot, we took for each concept the highest score over all shots.

A second approach made use of MorelikeThis Solr component (MLT) combined with Entityclassifier.eu annotation [1]. We created a temporary document from the query as the root for searching similar documents, and performed the search over segments from the LIMSI transcripts created using sliding windows and enriched with synonyms.

4. RESULTS

4.1 Search task

The results of the search task are listed in Table 1. We first notice than given the same conditions, subtitles perform significantly better than any of the transcripts, which is an expected outcome. It is also interesting to note that using the visual concepts in the query slightly increases the results for all measures (e.g., clustering10-C vs clustering10-noC).

Table 1: Results of the Search task

Run	MRR	mGAP	MASP
scenes-C	**0.3095**	**0.1770**	0.1951
scenes-noC	0.3091	0.1767	0.1947
scenes-S	**0.3152**	0.1635	**0.2021**
scenes-I	0.2613	0.1444	0.1582
scenes-U	0.2458	0.1344	0.1528
part-sc-C	0.2284	0.1241	0.1024
part-sc-noC	0.2281	0.1240	0.1021
cl10-C	0.2929	0.1525	0.1814
cl10-noC	0.2849	0.1479	0.1713
SW-60-S	0.2833	**0.1925**	**0.2027**
SW-60-I	0.1965	0.1206	0.1204
SW-40-U	0.2368	0.1342	0.1501

Overall, the best approaches are those using scenes and sliding windows. Scene based approaches retrieve a higher number of correct relevant segments within a time window of 60 seconds (higher MRR), but they are not the most precise in terms of start and end time, compared to the sliding windows approach (as suggested by mGAP and MASP).

4.2 Hyperlinking task

The results are listed in Table 2. For both LA and LC conditions, runs using scenes outperform other runs for all metrics. The MoreLikeThis/Entityclassifier.eu approach comes second. As expected, using the context increases the precision when hyperlinking video segments. It is also notable that the precision at rank n decreases when n increases.

Table 2: Results of the Hyperlinking task

Run	MAP	P-5	P-10	P-20
LA cl10	0.0577	0.4467	0.3200	0.2067
LA MLT	0.1201	0.4200	0.4200	0.3217
LA scenes	**0.1770**	**0.6867**	**0.5867**	**0.4167**
LC cl10	0.0823	0.5733	0.4833	0.2767
LC MLT	0.1820	0.5667	0.5667	0.4300
LC scenes	**0.2523**	**0.8133**	**0.7300**	**0.5283**

5. CONCLUSION

This paper presented our framework and results at the MediaEval Search and Hyperlinking task. From our runs, it is clear that scene segmentation is the approach with the best performances. Therefore, this approach should be studied more in depth, a potential improvement being to refine the segmentation using semantics or speakers information. Also, we see here that this task benefits from the use of visual information present in the video. Hence, those two axes should be the next steps to study for a future challenge.

6. ACKNOWLEDGMENTS

This work was supported by the European Commission under contracts FP7-287911 LinkedTV and FP7-318101 MediaMixer.

7. REFERENCES

[1] M. Dojchinovski and T. Kliegr. Entityclassifier.eu: Real-Time Classification of Entities in Text with Wikipedia. In H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *Lecture Notes in Computer Science*, pages 654–658. Springer Berlin Heidelberg, 2013.

[2] M. Eskevich, J. Gareth J.F., S. Chen, R. Aly, and R. Ordelman. The Search and Hyperlinking Task at MediaEval 2013. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[3] M. Eskevich, G. Jones, C. Wartena, M. Larson, R. Aly, T. Verschoor, and R. Ordelman. Comparing retrieval effectiveness of alternative content segmentation methods for Internet video search. In *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, pages 1–6, 2012.

[4] L. Lamel and J.-L. Gauvain. Speech processing for audio indexing. In *Advances in Natural Language Processing (LNCS 5221)*, pages 4–15. Springer, 2008.

[5] D. Lin. An Information-Theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[6] P. Sidiropoulos, V. Mezaris, and I. Kompatsiaris. Enhancing Video concept detection with the use of tomographs. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2013.

[7] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso. Temporal Video Segmentation to Scenes Using High-Level Audiovisual Features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(8):1163–1177, Aug. 2011.

[8] D. Stein, S. Eickeler, R. Bardeli, E. Apostolidis, V. Mezaris, and M. Müller. Think Before You Link – Meeting Content Constraints when Linking Television to the Web. In *Proc. NEM Summit*, Nantes, France, Oct. 2013. to appear.

[9] S. Tschöpel and D. Schneider. A lightweight keyword and tag-cloud retrieval algorithm for automatic speech recognition transcripts. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association ISCA (INTERSPEECH)*, 2010.