EURECOM @ SAVA2015: Visual Features for Multimedia Search

Maria Eskevich, Benoit Huet
EURECOM, Sophia Antipolis, France
maria.eskevich@gmail.com; benoit.huet@eurecom.fr

ABSTRACT

This paper describes our approach to carry out multimedia search by connecting the textual information, or the corresponding textual description of the required visual content, in the user query to the audio-visual content of the videos within the collection. The experiments were carried out on the dataset of the Search and Anchoring in Video Archives (SAVA) task at MediaEval 2015, consisting of roughly 2700 hours of the BBC TV broadcast material. We combined visual concepts extraction confidence scores with the information about corresponding word vectors distances in order to rerank the baseline text based search. The reranked runs did not outperform the baseline, however they exposed potential of our method for further improvement.

1. INTRODUCTION

Issuing a textual query for a search within a multimedia collection is a task that is familiar to the Internet users nowadays. The systems performing this search are usually based on the corresponding transcript content of the videos or on the available metadata. The link between the given textual description of the query, or of the required visual content, and the visual features that can be automatically extracted for all the videos in the collection has not been thoroughly investigated. In [13] the visual content was used to impose the segmentation units, while in [2] and [4] the visual concepts were used for reranking of the result list for the case of search performed for the hyperlinking task, i.e., video to video search. However, as the reliability of the extracted visual concepts and the types of the concepts themselves vary based on the training data and the task framework, it is still hard to transfer these systems output from one collection or task to another while keeping the same impact on improvement.

In this paper we describe our experiments that attempt to create this link between the visual/textual content of the query and the visual features of the collection by incorporating the information about the words vectors distance into the confidence scores calculation. We take into account not only the actual query words and words assigned to the visual concepts, but also their lexical context, calculated as close word vectors following the word2vec approach [10]. By expanding the list of terms for comparison by the lexical context, we attempt to deal with the potential mismatch of the terms used in the video and those describing the visual concepts, as the speakers in the videos might not directly describe the visual content, but it might be implied in the further lexical context of the topic of their speech.

We use the dataset of the Search and Anchoring task at MediaEval 2015 [5] that contains both textual and visual descriptions of the required content, thus we can compare the influence of words vectors similarity for the cases when we establish the connection between the textual query and the visual content within the collection, and between the textual description of the visual request and the visual content within the collection.

2. SYSTEM OVERVIEW

To compare the impact of our approach, we create a baseline run that all further implementations are based upon.

First, we divide all the videos in the collection into segments of a fixed length of 120 seconds with a 30 seconds overlap step. We store the corresponding LIMSI transcripts [8] as the documents collection, and the information about the start of the first word after a pause longer than 0.5 seconds or a first switch of speakers as the potential jump-in point for each segment, as in [6].

Second, we use the open-source Terrier 4.0. Information Retrieval platform[1] [11] with a standard language modeling implementation [7], with default *lamda* value equal to 0.15, for indexing and retrieval. The resulting top 1000 segments for each of the 30 queries represent the baseline result after the removal of the overlapping results.

Third, for these top 1000 segments we calculate a new confidence score that represents a combination of three values, see Equation 1: i) confidence score of the terms that are present both in the query, textual or visual field, ($C_{Q\_w_i}$) and in the visual concepts extracted for the segment ($C_{VC\_w_i}$); ii) confidence score of the terms that are present both in the query, textual or visual field, ($C_{Q\_w_i}$) and in lexical context of the visual concepts extracted for the segment ($C_{W2V4VC\_w_i}$); iii) confidence score of the terms that are present both in the lexical context of the query, textual or visual field, ($C_{W2V4Q\_w_i}$) and in the visual concepts extracted for the segment ($C_{VC\_w_i}$). We empirically chose to assign higher value (0.6) to the confidence score of the first type, as those are the words used in the transcripts and visual concepts, and lower equal values (0.2) for the scores using the lexical context, see Equation 1. We use the open-source implementation of the word2vec algorithm [2]

[1] http://www.terrier.org
[2] http://word2vec.googlecode.com/svn/trunk/

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

Table 1: Precision at ranks 5, 10, 20.

Query fields used	Visual concepts	P@5 overlap	P@5 bin	P@5 tol	P@10 overlap	P@10 bin	P@10 tol	P@20 overlap	P@20 bin	P@20 tol
text	none	0.6733	0.6400	0.6133	0.6133	0.5933	0.5467	0.4067	0.3983	0.3133
text	Oxford	0.4533	0.4467	0.400	0.4233	0.4167	0.3767	0.3133	0.3367	0.2667
visual	Oxford	0.4933	0.5000	0.4733	0.4633	0.4900	0.4333	0.3367	0.3683	0.2917
text	Leuven	0.4667	0.4333	0.4400	0.4567	0.4500	0.4300	0.3450	0.3667	0.3017
visual	Leuven	0.4400	0.4533	0.4000	0.4500	0.4333	0.4200	0.3500	0.3667	0.2883
text	CERTH	0.3600	0.3467	0.3400	0.3333	0.3467	0.3200	0.2450	0.2567	0.2167
visual	CERTH	0.3733	0.3600	0.3400	0.4133	0.3900	0.3933	0.2933	0.3050	0.2600

Table 2: Official metrics for all the runs

Query fields used	Visual concepts	MAP	MAP_bin	MAP_tol	MAiSP
text	none	0.5511	0.3529	0.3089	0.3431
text	Oxford	0.3196	0.2739	0.2053	0.2978
visual	Oxford	0.3368	0.2958	0.2293	0.3092
text	Leuven	0.3227	0.2801	0.2187	0.2958
visual	Leuven	0.3394	0.2970	0.2222	0.3117
text	CERTH	0.2295	0.2027	0.1554	0.1983
visual	CERTH	0.2624	0.2375	0.1822	0.2380

with the pre-trained vectors trained on part of Google News dataset [3] (about 100 billion words), cf. [9]. We take the top 100 word2vec output for consideration, remove the stop words from both the query and the word2vec output, and run Porter Stemmer [12] on all lists for normalization.

Finally, the new confidence score values are used for the reranking of the initial results, these are filtered for the overlapping segments, and the jump-in points of the segments are used as start times.

$$
\begin{aligned}
ConfScore = {} & \frac{\sum_{i=1}^{N_{Q\_VC}} (C_{Q\_w_i} * C_{VC\_w_i})}{N_{Q\_VC}} * 0.6+ \\
& + \frac{\sum_{i=1}^{N_{Q\_W2V4VC}} (C_{Q\_w_i} * C_{W2V4VC\_w_i})}{N_{Q\_W2V4VC}} * 0.2+ \\
& + \frac{\sum_{i=1}^{N_{W2V4Q\_VC}} (C_{W2V4Q\_w_i} * C_{VC\_w_i})}{N_{W2V4Q\_VC}} * 0.2
\end{aligned} \tag{1}
$$

3. EXPERIMENTAL RESULTS

Tables 1-2 show the evaluation results of the submissions. In both tables each line represent the an approach that used textual or visual query field (first column) and visual concepts extracted by Oxford [3], Leuven [14] or CERTH [1] systems. Although none of these runs outperforms the baseline, some trends can be tracked. According to all of the metrics in Table 2 the runs that use the connection between the visual query field and the visual concepts extracted for the collection achieve higher scores than the runs using the textual fields. This means that at least partly these visual concepts defined for the other task and extracted for this collection can be transferred to be used in this task. In terms of precision, the trends is not as consistent, as only the runs that use the Oxford and CERTH visual concepts have better scores when the visual query description is used for all the measurements, and the results based on the Leuven visual concepts extraction vary between different measurements.

[3]https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit

4. CONCLUSION AND FUTURE WORK

In this paper we have described a new approach to combine the confidence scores of the visual concepts extraction and the textual description of the query, weighted by the closeness of the terms in the words vector space.

Even though as expected we achieve higher scores for the runs using the closeness between the visual descriptions of the queries and the visual concepts, we achieve comparable results when using the textual descriptions. Therefore we envisage that further tuning of the confidence scores combination and reranking strategies can bring the results to the level of baseline and further improvement.

5. ACKNOWLEDGMENTS

This work was supported by the European Commission's 7th Framework Programme (FP7) under FP7-ICT 287911 (LinkedTV); Bpifrance within the NexGen-TV Project, under grant number F1504054U.

6. REFERENCES

[1] E. E. Apostolidis, V. Mezaris, M. Sahuguet, B. Huet, B. Cervenková, D. Stein, S. Eickeler, J. L. R. García, R. Troncy, and L. Pikora. Automatic fine-grained hyperlinking of videos within a closed collection using scene segmentation. In *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, pages 1033–1036, 2014.

[2] C. Bhatt, N. Pappas, M. Habibi, and A. Popescu-Belis. Idiap at MediaEval 2013: Search and Hyperlinking Task. In *MediaEval 2013 Workshop*, 2013.

[3] K. Chatfield and A. Zisserman. Visor: Towards on-the-fly large-scale object category retrieval. In *Computer Vision–ACCV 2012*, pages 432–446. Springer, 2013.

[4] S. Chen, M. Eskevich, G. J. F. Jones, and N. E. O'Connor. An investigation into feature effectiveness for multimedia hyperlinking. In *MultiMedia Modeling* -

20th Anniversary International Conference, MMM 2014, Dublin, Ireland, January 6-10, 2014, Proceedings, Part II, pages 251–262, 2014.

[5] M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. F. Jones. SAVA at mediaeval 2015: Search and anchoring in video archives. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, Wurzen, Germany, 2015.

[6] M. Eskevich and G. J. F. Jones. Time-based segmentation and use of jump-in points in DCU search runs at the search and hyperlinking task at mediaeval 2013. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 18-19, 2013.*, 2013.

[7] D. Hiemstra. *Using language models for information retrieval*. PhD thesis, University of Twente, The Netherlands, 2001.

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

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119, 2013.

[10] T. Mikolov, W. tau Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*, May 2013.

[11] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[12] M. F. Porter. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137, 1980.

[13] M. Sahuguet, B. Huet, B. Cervenková, E. E. Apostolidis, V. Mezaris, D. Stein, S. Eickeler, J. L. R. García, and L. Pikora. Linkedtv at mediaeval 2013 search and hyperlinking task. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 18-19, 2013.*, 2013.

[14] T. Tommasi, T. Tuytelaars, and B. Caputo. A testbed for cross-dataset analysis. *CoRR*, abs/1402.5923, 2014.