SAVA at MediaEval 2015: Search and Anchoring in Video Archives

Maria Eskevich[1], Robin Aly[2], Roeland Ordelman[2], David N. Racca[3], Shu Chen[3], Gareth J.F. Jones[3]

[1]EURECOM, Sophia Antipolis, France; [2]University of Twente, The Netherlands
[3]ADAPT Centre, School of Computing, Dublin City University, Ireland
maria.eskevich@gmail.com; {r.aly, ordelman}@ewi.utwente.nl;
{dracca, gjones}@computing.dcu.ie; shu.chen4@mail.dcu.ie

ABSTRACT

The Search and Anchoring in Video Archives (SAVA) task at MediaEval 2015 consists of two sub-tasks: (i) search for multimedia content within a video archive using multimodal queries referring to information contained in the audio and visual streams/content, and (ii) automatic selection of video segments within a list of videos that can be used as anchors for further hyperlinking within the archive. The task used a collection of roughly 2700 hours of the BBC broadcast TV material for the former sub-task, and about 70 files taken from this collection for the latter sub-task. The search sub-task is based on an ad-hoc retrieval scenario, and is evaluated using a pooling procedure across participants submissions with crowdsourcing relevance assessment using Amazon Mechanical Turk (MTurk). The evaluation used metrics that are variations of MAP adjusted for this task. For the anchor selection sub-task overlapping regions of interest across participants submissions were assessed using MTurk workers, and mean reciprocal rank (MRR), precision and recall were calculated for evaluation.

1. INTRODUCTION

Current developments in the technologies for recording and storing of multimedia content are leading to very rapid growth in the resulting multimedia archives. Moreover the digitisation of the content created in previous decades is being added to this contemporary material. This stored information can potentially be used by a wide variety of users including multimedia professionals, e.g. archivists, journalists, and the general public. We envisage the main aim of the SAVA task in assisting these different users in their interaction with the available collections by facilitating efficient access to relevant content. The solutions to the challenges of the SAVA task should help the users: 1) to retrieve interesting parts of the archived multimedia documents when issuing audio-visual queries to a search system; 2) to improve the browsing aspect of this activity by providing users with the content that has pre-defined or changing on-the-fly anchor points that can lead them to further discoveries on topics of interest within the collection. Thus the SAVA task consists of two sub-tasks: Search for multimedia content and Automatic anchor selection.

- **Search for multimedia content** This promotes the development of search methods that use multiple modalities (e.g., speech, visual content, speaker emotions, etc) to answer search queries by returning relevant video segments of unrestricted size. Similar to the earlier MediaEval 2013 Search & Hyperlinking edition of this sub-task [4], participants were provided with a two-fielded query, where one field refers to spoken content and the other refers to the visual content of relevant segments. Participants could use either or both fields to find video segments within the collection.

- **Automatic anchor selection** This explores methods to automatically identify anchors for a given set of videos, where anchors are media fragments (with their boundaries defined by their start and end time) for which users could require additional information. What constitutes an anchor depends on the video, e.g., in a news programme it could be a mention of persons, and in a documentary it could be the view of particular buildings. Participants were provided with a number of videos of different types and were requested to automatically identify anchors within these videos.

2. EXPERIMENTAL DATASET

The dataset for both sub-tasks is a collection of 4021 hours of videos provided by the BBC, which are split into a development set of 1335 hours, and a test set of 2686 hours. The average length of a video was roughly 45 minutes, and most videos were in the English language. The test collection was broadcast content of date spans 01.04.2008 – 11.05.2008 and 12.05.2008 – 31.07.2008 for the development and test sets respectively. The BBC kindly provided human generated textual metadata and manual transcripts for each video. Participants were also provided with the output of several content analysis methods, which we describe in the following subsections.

Although both sub-tasks are based on the same collection, they use different set of videos within each sub-task framework. For both development and testing of the system within the 'Search for multimedia content' sub-task the participants used the test set of the video collection. While the videos for the 'Automatic anchor selection' were taken from both development and test set of the video collection in order to have a uniform representation of the files containing previously defined manually created anchors that were used for sub-task assessment.

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

2.1 Audio Content

The audio was extracted from the video stream using the *ffmpeg* software toolbox (sample rate = 16,000Hz, no. of channels = 1). Based on this data, the transcripts were created using the following ASR approaches and provided to participants:

- LIMSI-CNRS/Vocapia[1] using the VoxSigma vrbs_trans system (version eng-usa_4.0) [7].
- The LIUM system[2] [11], is based on the CMU Sphinx project. The LIUM system provided three output formats: (1) one-best transcripts in NIST CTM format, (2) word lattices in SLF (HTK) format, following a 4-gram topology, and (3) confusion networks in a format similar to ATT FSM.
- The NST/Sheffield system[3] is trained on multi-genre sets of BBC data that do not overlap with the collection used for the task, and uses deep neural networks [8]. The ASR transcript contains speaker diarization, similar to the LIMSI-CNRS/Vocapia transcipts.

Additionally, prosodic features were extracted using the OpenSMILE tool version 2.0 rc1 [6][4]. The following list of prosodic features were calculated over sliding windows of 10 milliseconds: root mean squared (RMS) energy, loudness, probability of voicing, fundamental frequency (F0), harmonics to noise ratio (HNR), voice quality, and pitch direction (classes falling, flat, raising, and direction score).

2.2 Visual Content

The computer vision groups at University of Leuven (KUL) and University of Oxford (OXU) provided the output of concept detectors for 1,537 concepts from ImageNet[5] using different training approaches. The approach by KUL uses examples from ImageNet as positive examples [12], while OXU uses an on-the-fly concept detection approach, which downloads training examples through Google image search [3].

3. TASK INPUT DEFINITION

As we assumed that both types of user activities behind the sub-tasks frameworks can be carried out by both professionals and general audience, we involved representatives of both user categories into the ground truth creation:

- **Search for multimedia content:** 9 development set and 30 test set queries were defined by professionals with the following profile: 1) they work in the field, e.g. they were journalists, archivists, etc; 2) they were native English speakers, and 3) they were generally familiar with BBC content. For each query in the development set these users defined two relevant video segments in order to ensure the existence of potential relevant content for an ad hoc search.

- **Automatic anchor selection:** We used the video files containing the manually defined anchors in 2013-2014 Search & Hyperlinking tasks [4, 5]: 42 and 33 files respectively for the development and testing of the approaches. The users represented the general public: they had to be 18-30 years old and had to use search engines and services such as *Youtube* on a daily basis. The anchors provided in this ground truth are by no means exhaustive, they only exemplify potential anchors that can be defined within a given video.

More elaborate description of this user study design and the anchor definition procedure can be found in [2] and [9] respectively.

4. REQUIRED RUNS

As our evaluation makes use of cross-comparison between runs, we did not limit the participants in the number of submissions for either of the tasks. However, we stated that due to finite resources, only limited number of runs would be assessed through crowdsourcing.

5. RELEVANCE ASSESSMENT AND EVALUATION METRICS

To evaluate the submissions of the search sub-task, First, the runs were normalised: videos with corrupted audio-visual content due to bugs in the employed software ffmpeg were dismissed, segments shorter than 10 seconds were expanded to this length, segments longer than 2 minutes were cut after this length (using the original's segment start), and segments overlapping with previously returned segments were adjusted to remove the overlap. Second, we used the pooling method with selected runs. Third, the top 10 ranks of all submitted runs were evaluated using crowdsourcing technologies. We report precision oriented metrics, such as precision at various cutoffs and mean average precision (MAP), using different approaches to take into account segment overlap, as described in [1, 10].

For the anchoring sub-task, we used the top-25 ranks of all submissions, and merged overlapping segments. The resulting segments were judged by MTurk workers who gave their opinion on these segments taken from the context of the videos. For the MRR, recall/precision, a result segment in the run is judged relevant if it overlaps with a relevant combined segment.

6. SUMMARY AND CONCLUSIONS

This paper describes the setup of the search and anchoring sub-tasks at the MediaEval 2015. While the definition of the search task is built on the experience of several years, the anchoring sub-task was new in 2015. Here, we describe the data provided to the task participants and the methods used to generate the input data and to evaluate submitted results.

7. ACKNOWLEDGMENTS

This work was supported by the European Commission's 7th Framework Programme (FP7) under FP7-ICT 269980 (AXES) and FP7-ICT 287911 (LinkedTV); Bpifrance within the NexGen-TV Project, under grant number F1504054U; the Dutch national program COMMIT/; Science Foundation Ireland (Grant No 12/CE/I2267) as part of the Centre for Next Generation Localisation (CNGL) project at DCU. The user studies were executed in collaboration with Jana Eggink and Andy O'Dwyer from BBC Research, to whom the authors are grateful.

[1] http://www.vocapia.com/

[2] http://www-lium.univ-lemans.fr/en/content/language-and-speech-technology-lst

[3] http://www.natural-speech-technology.org

[4] http://opensmile.sourceforge.net/

[5] http://image-net.org/popularity_percentile_readme.html

8. REFERENCES

[1] R. Aly, M. Eskevich, R. Ordelman, and G. J. F. Jones. Adapting binary information retrieval evaluation metrics for segment-based retrieval tasks. Technical Report 1312.1913, ArXiv e-prints, 2013.

[2] R. Aly, R. Ordelman, M. Eskevich, G. J. F. Jones, and S. Chen. Linking inside a video collection - what and how to measure? In *Proceedings of the 22nd International Conference on World Wide Web Companion, IW3C2 2013, Rio de Janeiro, Brazil*, pages 457–460, Brazil, May 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] M. Eskevich, R. Aly, R. Ordelman, S. Chen, and G. J. F. Jones. The Search and Hyperlinking Task at MediaEval 2013. In *MediaEval 2013 Workshop*, Barcelona, Spain, 2013.

[5] M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. F. Jones. The Search and Hyperlinking task at MediaEval 2014. In *Working Notes Proceedings of the MediaEval 2014 Workshop*, Barcelona, Catalunya, Spain, 2014.

[6] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of ACM Multimedia 2013*, pages 835–838, Barcelona, Spain.

[7] J.-L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News transcription system. *Speech Communication*, 37(1-2):89–108, 2002.

[8] P. Lanchantin, P. Bell, M. J. F. Gales, T. Hain, X. Liu, Y. Long, J. Quinnell, S. Renals, O. Saz, M. S. Seigel, P. Swietojanski, and P. C. Woodland. Automatic transcription of multi-genre media archives. In *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM@INTERSPEECH)*, volume 1012 of *CEUR Workshop Proceedings*, pages 26–31. CEUR-WS.org, 2013.

[9] R. J. F. Ordelman, M. Eskevich, R. Aly, B. Huet, and G. J. F. Jones. Defining and evaluating video hyperlinking for navigating multimedia archives. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy - Companion Volume*, pages 727–732, 2015.

[10] D. N. Racca and G. J. F. Jones. Evaluating Search and Hyperlinking: an example of the design, test, refine cycle for metric development. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, Wurzen, Germany, 2015.

[11] A. Rousseau, P. Deléglise, and Y. Estève. Enhancing the TED-LIUM corpus with selected data for language modeling and more ted talks. In *The 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, May 2014.

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