A Decade of Work on Semantic Concept Detection from Video: TRECVid's Contribution

Alan F. Smeaton[1], Georges Quénot[2], and Wessel Kraaij[3]

[1]CLARITY: Centre for Sensor Web Technologies, School of Computing,
Dublin City University, Glasnevin, Dublin 9, Ireland.
[2]UJF-Grenoble 1/UPMF-Grenoble 2/Grenoble INP, CNRS,
LIG UMR 5217, 38041 Grenoble, France.
[3]Institute for Computing and Information Sciences Radboud University Nijmegen
and TNO ICT, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands.
alan.smeaton@dcu.ie

We are interested in content-based tasks for visual media, especially video. Automatic assignment of semantic tags representing visual or multimodal concepts ("high-level features") to video segments is a fundamental technology for filtering, categorization, browsing, search, and other video exploitation activities. Approaches based on using metadata, automatically recognised speech, image-image similarity, and object detection and matching all have their contributions but using semantic concepts is the one which is most scalable and has an interesting intersection with language.

TRECVid is an annual benchmarking activity which has been on-going for 10 years [1]. It has focused on many tasks such as shot boundary detection, summarisation, and various forms of search. The 2012 running of TRECVid addressed 6 tasks, including Semantic INdexing (SIN) of video.

The rationale for the SIN task is as follows. Given a test collection of video, a master shot reference, and a set of concept definitions, return for each concept a list of at most 2,000 shot IDs from the test collection ranked according to their likelihood of containing the concept. The data set is 291 hours of short videos with durations between 10 seconds and 3.5 minutes where there exists at least 4 positive samples of each concept. Of the 346 test concepts, a subset of 50, not known by participants at submission time, are manually judged.

This task has run since the start of TRECVid and in 2012, the task is similar to previous years, except in scale. Previously, for the majority of participants the approach was to build independent concept detectors based on training a machine learning toolkit on a set of positive and negative example shots, using low-level features from the images, as described in [2]. To cater for this, a collaborative annotation of the concepts is carried out each year to build a concept bank on which to train classifiers. The TRECVid organizers also provide a set of relations between the concepts of two types: A implies B and A excludes B. Relations that can be derived by transitivity are not included. Participants are free to use the relations or not and submissions are not required to comply with them. Some of the more advanced methods from participants use the annotations of non-evaluated concepts and the ontology relations to improve the detection of the evaluated concepts.

The ontological relations among the concepts and an active learning method are used in the collaborative annotation to ensure that each annotation made is as useful as possible and especially that as many positive samples of the sparse concepts as possible are obtained while the negative ones are as close as possible to the class boundary. This helps to optimize the annotation effort. Participation in the collaborative annotation is encouraged but it is not required to have access to the full annotation.

This year in TRECvid we introduced a task on indexing of concepts in video using only data from the web or from archives without the need of additional annotations, and the results of this point to a scalable roll-out of concept detection without the need for manually annotated training data. There was also considerable progress in the creation of language-independent "virtual" concepts derived from query topics at search time. TRECVid also introduced a "concept pair" task requiring the detection of pairs of unrelated concepts such as "Beach + Mountain"; "Old_People + Flags"; "Animal + Snow", etc. instead of the independent detection of simple concepts. This was introduced to promote the development of methods for retrieving shots containing a combination of concepts that do better than just combining the output of individual concept detectors.

TRECVid is an important contributor in the area of multimedia information retrieval. Over the last decade it has involved over 1,200 researchers from hundreds of research groups worldwide and its impact on the area of multimedia IR is well-documented [3]. The scale of TRECVid participants' achievements in terms of the sizes of datasets and numbers of concepts being detected, means that we now offer a serious opportunity for collaboration with other research areas, at scale. In this poster/presentation we provide further details on the results from the TRECVid 2012 workshop (held the last week November 2012) as well as an overview of the techniques used.

Acknowledgements: AS thanks Science Foundation Ireland (grant 07/CE/I147)

References

1. A.F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM international workshop on Multimedia Information Retrieval*, pages 321–330. ACM, 2006.
2. A.F. Smeaton, P. Over, and W. Kraaij. High level feature detection from video in TRECVid: a 5-year retrospective of achievements. In *In Ajay Divakaran (Ed.), Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer US, Norwell, MA, 2008.
3. C.V. Thornley, A.C. Johnson, A.F. Smeaton, and H. Lee. The scholarly impact of TRECVid (2003–2009). *Journal of the American Society for Information Science and Technology*, 62(4):613–627, 2011.