

Short Text Queries for Video Retrieval

Multimedia Event Detection at TRECVID 2013

Team UMass

James Allan, Jeff Dalton, John Foley, R. Manmatha, Venkatesh Murthy, David Wemhoener
{allan, jdalton, jfoley, manmatha, venk, wem} @cs.umass.edu

Center for Intelligent Information Retrieval
School of Computer Science
University of Massachusetts Amherst

Abstract

UMass submitted runs to TRECVID 2013 that focus on zero exemplar (0Ex) performance for both the pre-specified (PS) and ad-hoc (AH) events. We participated on our own and in cooperation with our partners, led by SRI-AURORA. While our partners focused on video and image features and using exemplars in event detection, we focused for those runs on OCR and ASR as well as 0Ex performance. In contrast, the UMass submissions explored what happens if we keep the queries short, allowing only a handful of words as input to our system. Difficulties inherent to the task include mapping the text input to video and image features, being unable to use statistics from the target collection, and having limited training data, in terms of number of queries and number of videos. In spite of these challenges, UMass was able to achieve strong results on OCR and ASR based runs, and do second-best overall for the PS events. In addition, our performance relative to that of our partner's indicates that shorter, more realistic queries have a potential for achieving good results.

Description of Submitted Runs

Pre-Specified Task:

UMass_MED13_VisualSys_PROGAIL_PS_0Ex_1: This run uses only our short query to identify concepts. The top few selected concepts are then used with a Query Likelihood model. The weight of a visual concept, c , in a video, v , is summarized using the approximation:

$$(\text{mean}(c, v) + \text{stddev}(c, v)) * 100$$

UMass_MED13_OCRSys_PROGAIL_PS_0Ex_1: This run uses only our short query matched against text extracted from OCR on the video frames.

UMass_MED13_ASRSys_PROGAIL_PS_0Ex_1: This run uses only our short query matched against text extracted from ASR on the audio track.

UMass_MED13_FullSys_PROGAIL_PS_0Ex_1: This run combines the aforementioned runs, using our short query to match visual concepts as well as OCR and ASR features and the

results are combined using a trained linear model.

UMass_MED13_FullSys_PROGAll_PS_100Ex_1: This run uses the full query, matched against ASR, OCR, and visual concepts. In addition, information from pseudo annotations and set-based text metrics are included based on the exemplar videos.

AdHoc Task:

Our participation in the Ad-Hoc events mirrors our submissions for the Pre-Specified events. Those run identifiers are listed below; see the corresponding runs above for a description what was done.

UMass_MED13_OCRSys_PROGAll_AH_0Ex_1
UMass_MED13_ASRSys_PROGAll_AH_0Ex_1
UMass_MED13_VisualSys_PROGAll_AH_0Ex_1
UMass_MED13_FullSys_PROGAll_AH_0Ex_1
UMass_MED13_FullSys_PROGAll_AH_100Ex_1

Introduction

Multimedia is available on the web in a scale that prevents realistic human annotation. In order to access and search through this data automated methods of event detection are needed. Pushing this technology forward is a motivation for participation in the Multimedia Event Detection task of TRECVID (Over et al. 2013). Extraction, indexing and search of low level, high level, and semantic features are common across many systems, and performance reported in previous TRECVID competitions is strong. In recent years we have become interested in what we call “Zero-Shot” retrieval, where there are no exemplar videos, but where the task remains the same: to find videos where an event has occurred.

The University of Massachusetts participated in TRECVID 2013, with our partners led by SRI-AURORA. For that group, we developed techniques intended for the zero exemplar (0ex) runs. In particular, we explored methods for selecting video concepts that matched the textual description of an event and for retrieving videos by those concepts. We also investigated improvements caused by manually editing the event descriptions into queries based on knowledge of the underlying retrieval model. For a full description of that run, see the SRI-AURORA paper (Liu et al. 2014).

One of our interests in this task is what might be called “realistic search” where the user has no sample videos at hand and only a poor description of their information need. We set out to explore that question on our own, leveraging the SRI-AURORA team’s video annotation but changing our query processing. This notebook paper sketches our efforts in that direction and our findings to date.

Overview of process

We used video concepts, OCR output, and ASR output generated by our partners in the SRI-AURORA team. We had 847 concept detectors available to us, and for each we built a language model from the top 100 web pages returned from the ClueWeb '09B dataset with spam filtering. Using these language models, we were able to take a text query and retrieve relevant video, image, and action concepts. The text query was processed against those language models using the state-of-the-art sequential dependence model (Metzler and Croft 2005). The same retrieval model was used for the OCR and ASR text.

Generating final scores involved a building a linear fusion model across types of video concepts for the “visual” run, and for all different sources (visual as well as OCR and ASR) for the “full” run.

For more more details on data generation and search, see the 0Ex section of the SRI AURORA paper (Liu et al. 2014).

Generating extremely short queries

Using three human annotators, we created and adjudicated text queries based on the original “name” of the event-kit description. Most queries represent something simple a user would type in, like “birthday party” for event 6, reflecting what the annotators believed they would type into a commercial search engine in order to find web pages on the topic. Other events were a little more broad than the average web query -- for example, event 9 “Getting a vehicle unstuck” was changed to “car stuck snow mud sand” to better represent a set of events in which vehicles need to be freed from various materials, and the annotators settled on a more broad representation of the topic. After the annotators created their lists, they worked as a group to select a single query interpretation of the event.

It should be noted that our annotators were all researchers in Information Retrieval so they ought to be considered “expert” searchers. However, the resulting queries (below) seem similar to what we would expect from non-experts.

On average, these queries were much shorter than the “full” queries we had developed or that we had automatically extracted from the event description. Specifically, the full queries had an average of about 340 words per event, while these short queries had only 3.0 words per event on average.

The following table lists the event queries that were used for all UMass submissions. We highlight that this was the *entire* input to the search system for each event: there were no exemplars, no video concepts were manually identified, no words or phrases were highlighted,

nothing was marked as significant for OCR or ASR, no weighting was included, no synonyms were provided, and so on.

Evt #	Query	Evt #	Query	Evt #	Query
6	birthday party	21	bike trick bike stunts	31	honey bee beekeeping apiculture
7	changing flat tire	22	appliance cleaning	32	wedding shower bridal shower
8	flash mob	23	dog show	33	bike repair skateboard repair sailboat
9	car stuck snow mud sand	24	giving directions	34	musical instrument repair
10	animal grooming	25	marriage proposal	35	horseback riding competition horse riding equestrian competition
11	sandwich preparation	26	fixing home	36	felling tree cutting down tree chopping
12	parade	27	rock climbing	37	parking car parking truck
13	parkour	28	town hall government meeting	38	fetch stick ball dog
14	repairing appliance	29	win race	39	tailgate party grilling
15	sewing	30	metal working	40	tuning music instrument guitar piano violin

Required 100ex run

As a requirement of participation, we submitted in the 100-exemplar category as well. This run is not related to our “short query, no exemplars” research interest, so in addition to the features used in our zero-exemplar runs, we also employed some exemplar-based video and text features. We also used the full queries for this run, hoping to give our submission the best chance of success.

For the video features, we included event-specific classifiers based on “pseudo-annotations.” These annotations are built automatically from images and are represented as a distribution

across 1000 “concepts” (though without attached semantics) in various locations of the video frame (Can and Manmatha 2012). These were the only non-semantic video feature utilized in our full run.

In addition to the pseudo-annotations, we built a learning-to-rank style model based on text similarity features extracted from the exemplars.

While our performance was reasonable, it was not amongst the leaders and we hope to include more low-level features in future systems.

Results and Discussion

The discussion of results below reflects the final scores in terms of mean average precision (MAP) on the full PROGAll dataset.

Pre-specified queries

The final results for the Pre-Specified events for UMass, using our shorter, hopefully more realistic queries are shown in the table below. For comparison, we also include the corresponding numbers from the SRI-AURORA run that used the same features, except that it started from the full 340-word queries (Liu et al. 2014).

ex0 runs	OCR	ASR	Visual	Full
UMass (short queries)	3.3%	2.3%	5.1%	5.6%
SRI-AURORA (full queries)	3.7%	3.0%	6.5%	7.0%

While the full query representation used by our partner surely led to better results, the differences were not as significant as we would have expected given that the shorter queries were two orders of magnitude smaller. We hope to incorporate a round of feedback next year, using the short queries to expand based on terms found in a knowledge base such as Wikipedia, in hopes of reclaiming some of the lost performance.

Ad-hoc queries

The pre-adjudicated MED system numbers on the Ad-Hoc events are displayed below. Again, we show both our own runs and, for comparison, the larger-query runs from SRI-AURORA.

ex0 runs	OCR	ASR	Visual	Full
UMass (short queries)	3.9%	2.1%	0.5%	1.0%
SRI-AURORA (full queries)	4.3%	3.9%	0.6%	1.4%

While our OCR and ASR performance was still strong, performance was extremely weak on visual features compared to our performance on the Pre-Specified events, and especially in relation to other teams.

Naturally, the weaker visual system, as a component of our full system, resulted in the low score in "Full". The problem appears to be the result of a mismatch between some visual detector's description and what they actually find. This problem is more pronounced in the ad-hoc setting because the detectors were not designed for these queries. We are continuing to investigate this issue and are exploring possible solutions.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11-PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

Data Sources

ClueWeb09 - <http://lemurproject.org/clueweb09/> We use Category B filtered to documents with a spam score of less than 60. See <http://plg.uwaterloo.ca/~gvcormac/clueweb09spam/>

Galago Search Engine - <http://lemurproject.org/galago.php> - Our search was powered by the open-source Galago Search Engine, written in Java for the Lemur Project.

References

Can, E. and Manmatha, R., "A New Representation of Videos Using Pseudo-Annotations for Event Detection," CIIR Technical Report 2012.

Dalton, J., Mirajkar, P. and Allan, J., "Zero-Shot Video Retrieval Using Content and Concepts," a poster presentation at CIKM 2013, San Francisco, October 27-28, 2013.

Metzler, Donald, and W. Bruce Croft. "A Markov random field model for term dependencies." In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 472-479. ACM, 2005.

Liu, J., Cheng, H., Javed, O., Yu, Q., Chakraborty, I., Zhang, W., Divakaran, A., Sawhney, H., Allan, J., Manmatha, R., Foley, J., Shah, M., Dehghan, A., Witbrock, M., Curtis, J., Friedland, G. "SRI-Sarnoff AURORA System at TRECVID 2013 Multimedia Event Detection and Recounting"

TRECVID 2013.

Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Wessel Kraaij, Alan F. Smeaton, and Georges Quenot. Trecvid 2013 - an overview of the goals, tasks, data, evaluation mechanisms and metrics. In Proceedings of TRECVID 2013. NIST, USA, 2013.

Smeaton, A. F., Over, P., and Kraaij, W. 2006. Evaluation campaigns and TRECVID. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, New York, NY, 321-330. DOI= <http://doi.acm.org/10.1145/1178677.1178722>

Spitkovsky, Valentin I., and Angel X. Chang. "A Cross-Lingual Dictionary for English Wikipedia Concepts." In LREC, pp. 3168-3175. 2012.