Revision of LSCOM Event/Activity Annotations

DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia
Columbia University ADVENT Technical Report #221-2006-7, December 2006

Event/Activity Annotation Lead:

- Lyndon Kennedy, Columbia University, lyndon@ee.columbia.edu

Workshop Co-PIs:

- Milind Naphade, IBM T.J. Watson Research, naphade@us.ibm.com
- Alex Hauptmann, Carnegie Mellon University, alex@cs.cmu.edu
- John R. Smith, IBM T.J. Watson Research, jsmith@us.ibm.com
- Shih-Fu Chang, Columbia University, sfchang@ee.columbia.edu

Summary

The DTO-sponsored LSCOM workshop has developed a large concept lexicon for multimedia which includes approximately 1000 concepts related events, objects, locations, people, and programs [1, 3]. Human subjects annotated 449 of those concepts over a corpus of 80 hours of TRECVID 2005 videos [2], composed of 61,901 subshots, by visually inspecting a keyframe, which is a single still image, from each subshot and making a judgment as to whether a given concept is present or absent within that keyframe. This keyframe-only approach is fast and has proven to be sufficient for many of the static concepts, such as “airplane,” “car,” or “person;” however, event or activity concepts, like “airplane taking off,” “car driving,” or “person walking,” have a temporal component and are difficult for humans to judge, given only a single keyframe. It is necessary, therefore, to acquire labels based on a viewing of the video in motion to truly make a reliable annotation for the event. From the total 449 concepts that were annotated using the keyframe-based method, we selected 24 to be re-annotated using a video-based approach. These video-based annotations for 24 concepts are freely available for download, along with the keyframe-based labels for all 449 concepts [4].

To gather these video-based judgments, we designed a simple tool for providing binary positive/negative labels based on viewing a video clip of a subshot. The tool is web-based and can be used remotely with any modern web browser. It allows the user to work on only one subshot and concept at a time. The user can input the labels using a mouse or keyboard commands and is rapidly shown the next subshot to label as soon as a label is entered. A snapshot of the interface for the tool is shown in Figure 1.

Video-based annotation takes orders of magnitude longer than keyframe-based annotation, and we did not have the resources to conduct labeling over the entire corpus for even the small subset of event and activity concepts. To address this, we use the keyframe-based labels as a pre-filter for the video-based annotation. For example, to label the event concept "airplane taking off," we only had annotators look at subshots for which keyframe-based annotation had yielded positive labels for a few selected related concepts such as ("airplane," “airplane taking off,” “airplane landing,” “airplane flying”) and assumed that subshots which were negative for these concepts in the keyframe-based labels would also most likely still be negative for "airplane taking off." This cut the annotation task to only a few hundred subshots, down from 61,901, a considerable improvement. In the end, we select 24 concepts to be re-annotated using the video-based method, each over just a subset of the entire corpus, which resulted in a total of 37,450 video-based labels. This task took approximately 400 hours of labor to complete, or slightly more than 30 seconds per annotation.

Table 1 shows each of the 24 concepts that were re-annotated using the video-based method, along with the number of subshots that were examined for each concept. Since the video-based annotations are essentially

a revision of the keyframe-based labels, it is important to evaluate the difference in labels resulting from the two processes. Assuming that the video-based labels are more correct than the keyframe-based ones, it is interesting to see that for 78% (nearly 4 in 5) of the subshots examined, the video-based approach confirms that the keyframe-based label was already correct. The remaining subshots fall in to two categories, either: (1) the label is changed from "negative" to "positive," or (2) the label is changed from "positive" to "negative." We see that the second case is nearly twice as likely as the first. This is perhaps due to the design of the re-annotation process: we mostly pre-filter the video-based annotation task using subshots which were already judged to be "positive" by the keyframe-based approach, thereby making us more likely to catch "false positives" than "misses." We do, however, include some subshots for which the keyframe-based approach yielded positive labels for like-related concepts, so we can to some extent capture "misses" from the keyframe-based approach.

Concept Name	***# subshots***	***% same***	***% N→P***	***% P→N***
Airplane_Crash	574	99%	0%	0%
Airplane_Flying	570	83%	7%	10%
Airplane_Landing	570	94%	3%	3%
Airplane_Takeoff	570	95%	2%	4%
Car_Crash	4201	98%	1%	1%
Cheering	548	51%	0%	49%
Dancing	1027	42%	0%	58%
Demonstration_Or_Protest	2052	56%	32%	12%
Election_Campaign_Debate	497	95%	2%	3%
Election_Campaign_Greeting	497	66%	13%	21%
Exiting_Car	4201	92%	2%	6%
Fighter_Combat	743	74%	1%	25%
Greeting	394	80%	0%	20%
Handshaking	132	83%	0%	17%
Helicopter_Hovering	88	61%	30%	9%
People_Crying	138	44%	0%	56%
People_Marching	1937	39%	1%	60%
Riot	2052	93%	4%	4%
Running	6401	89%	9%	2%
Shooting	1483	77%	13%	10%
Singing	835	92%	0%	8%
Street_Battle	1483	48%	4%	47%
Throwing	56	46%	0%	54%
Walking	6401	68%	18%	14%
Total	**37450**	**78%**	**8%**	**14%**

Table 1. The 24 concepts event and activity concepts re-annotated by examining video keyframes, showing the total number of subshots re-annotated per concept (# subshots), the percentage of those subshots for which the annotation remained unchanged (% same), the percentage for which the annotation switched from "negative" to "positive" (%N→P), and the percentage for which the annotation switched from "positive" to "negative" (%P→N). Percentages may not add up to 100 due to rounding.

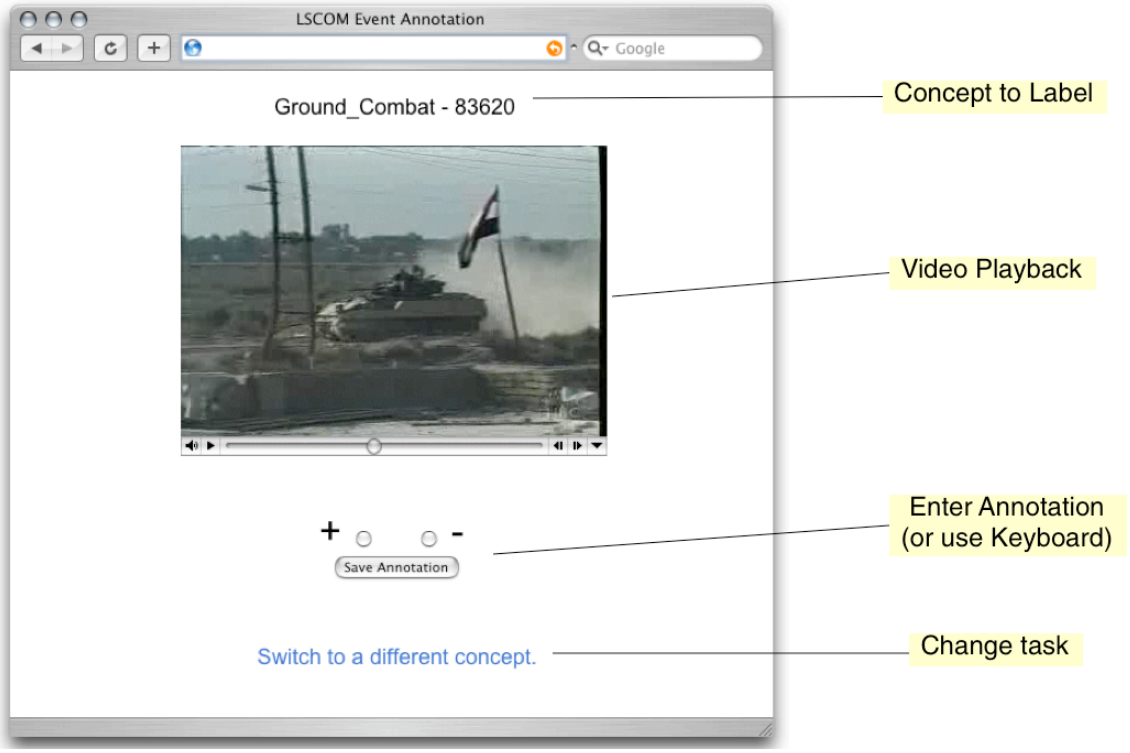

Figure 1. Interface used to gather video-based labels for event/activity concepts.

Acknowledgments

The LSCOM project has been sponsored by the Disruptive Technology Office (DTO). This material is based upon work funded in whole by the U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

References

[1] M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. Hauptmann, “A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005 (LSCOM-Lite),” IBM Research Technical Report, 2005.

[2] NIST TREC Video Retrieval Evaluation. http://www-nlpir.nist.gov/projects/trecvid/

[3] "LSCOM Lexicon Definitions and Annotations Version 1.0, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia." ADVENT Technical Report #217-2006-3 Columbia University, March 2006.

[4] LSCOM Lexicon Definitions and Annotations Download Site. http://www.ee.columbia.edu/dvmm/lscom.