Language-driven video retrieval

Andrei Barbu
Massachusetts Institute of Technology
andrei@0xab.com

N. Siddharth
Stanford University
nsid@stanford.edu

Jeffrey Mark Siskind
Purdue University
qobi@purdue.edu

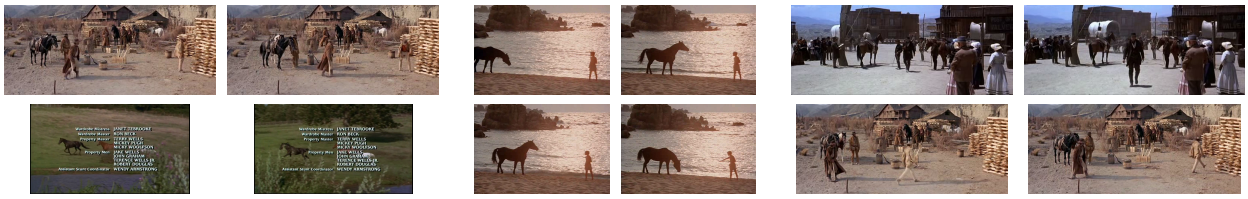
Abstract

We present an approach to searching large video corpora for video clips which depict a natural-language query in the form of a sentence. This approach uses compositional semantics to encode subtle meaning that is lost in other systems, such as the difference between two sentences which have identical words but entirely different meaning: The person rode the horse vs. The horse rode the person. *Given a video-sentence pair and a natural-language parser, along with a grammar that describes the space of sentential queries, we produce a score which indicates how well the video depicts the sentence. We produce such a score for each video clip in a corpus and return a ranked list of clips. Furthermore, this approach addresses two fundamental problems simultaneously: detecting and tracking objects, and recognizing whether those tracks depict the query. Because both tracking and object detection are unreliable, this uses knowledge about the intended sentential query to focus the tracker on the relevant participants and ensures that the resulting tracks are described by the sentential query. While earlier work was limited to single-word queries which correspond to either verbs or nouns, we show how one can search for complex queries which contain multiple phrases, such as prepositional phrases, and modifiers, such as adverbs.*

Video search engines lag behind text search engines in their wide use and performance. This is in part because the most attractive interface for finding videos remains a natural-language query in the form of a sentence but determining if a sentence describes a video remains a difficult task. This task is difficult for a number of different reasons: unreliable object detectors which are required to determine if nouns occur, unreliable event recognizers which are required to determine if verbs occur, the need to recognize other parts of speech such as adverbs or adjectives, and the need for a representation of the semantics of a sentence which can faithfully encode the desired natural-language query. We propose an approach which simultaneously addresses all of these problems [4]. Systems to date generally attempt to independently address the various aspects that make this task difficult. For example, they attempt to separately find videos that depict nouns and videos that depict verbs and essentially take the intersection of the two sets of videos [3]. This general approach of solving these problems piecemeal cannot represent crucial distinctions between otherwise similar input queries. For example, if you search for *The person rode the horse* and for *The horse rode the person*, existing systems would give the same result for both queries as they each contain the same words, but clearly the desired output for these two queries is very different. We develop a holistic approach which both combines tracking and word recognition to address the problems of unreliable object detectors and trackers and at the same time uses compositional semantics to construct the meaning of a sentence from the meaning of its words in order to make crucial but otherwise subtle distinctions between otherwise similar sentences.

In order to combine trackers and word recognizers to build task-specific models which recognize target sentences, we choose representations for trackers and word recognizers such as that they share the same underlying structure. Both employ a lattice and use the same inference algorithm, dynamic programming through the Viterbi [5] algorithm. Tracker lattices [1] are constructed from the output of an object detector [2], where detections in adjacent frames are connected to each other. Detections are weighted by their object detector score and the links between detections are weighted by the a measure of the optical flow [6] between pairs of detections in adjacent frames. The optimal path through this lattice represents an object track. Words are represented as finite state machines (FSMs) which accept tracks. The path through an FSM determines if one or more tracks participate in the event or relationship described by that word.

Given an input sentence and a single video we employ a dependency parser to extract the number of participants and the dependency structure which relates participants and words. For each participant we instantiate a tracker and for each word we instantiate an word recognizer. Rather than first tracking the participants and then checking if those tracks conform to the sentential query, the algorithm perform both steps simultaneously by taking the cross-product



(a) *The horse approached the person*

(b) *The person approached the horse*

Figure 1. The top 6 hits for two sentences from a corpus of 10 Hollywood movies. In both cases, half are true positives. The fact that the results are different shows that our method encodes the meaning of the entire sentence along with which object fills which role in that sentence.

of tracker and word recognizer lattices. This is made possible because both trackers and word recognizers share the same lattice structure and inference algorithm. The same dynamic programming algorithm is used to find the optimal path through this combined lattice for the sentential query. In order to encode the compositional semantics of the sentence we take only those cross-products which correspond to the dependency structure of the sentence. This allows our approach to discriminate between two sentences composed of the same words with different dependency structures. We call this algorithm, which produces a score for how well a sentence is depicted by a video along with a set of tracks for that sentence, the *sentence tracker*.

Given an input sentence and a set of videos we first split long videos into short overlapping clips in order to detect multiple events in each video. For each video clip in a corpus we employ the sentence tracker to determine if the video depicts the input sentence and to recover a set of tracks that for that sentence. Each video is scored by the quality of its tracks, which are guaranteed by construction to depict our target sentence, and the final score correlates with our confidence that the resulting tracks correspond to real objects in the video. We produce a score for every video-sentence pair and return multiple video hits ordered by their scores.

Our approach, unlike previous work, allows for a natural-language query of video corpora which have no human-provided annotation. This approach provides two novel video-search capabilities. First, it can encode the semantics of sentences compositionally, allowing it to express subtle distinctions such as the difference between *The person approached the horse* and *The horse approached the person*, Figure 1. Second, it can also encode structures more complex than just nouns and verbs, such as modifiers, *e.g*. adverbs, and entire phrases, *e.g*. prepositional phrases.

Acknowledgments

This research was supported, in part, by ARL, under Cooperative Agreement Number W911NF-10-2-0060, and the Center for Brains, Minds and Machines, funded by NSF STC award CCF-1231216. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either express or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein.

References

[1] A. Barbu, N. Siddharth, A. Michaux, and J. M. Siskind, "Simultaneous object detection, tracking, and event recognition," *Advances in Cognitive Systems*, vol. 2, pp. 203–220, 2012.

[2] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2241–2248.

[3] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 41, no. 6, pp. 797–819, 2011.

[4] N. Siddharth, A. Barbu, and J. M. Siskind, "Seeing what you're told: Sentence-guided activity recognition in video," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.

[5] A. J. Viterbi, "Convolutional codes and their performance in communication systems," *IEEE Transactions on Communication*, vol. 19, pp. 751–772, Oct. 1971.

[6] M. Werlberger, T. Pock, and H. Bischof, "Motion estimation with non-local total variation regularization," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2464–2471.