Integrating knowledge about affordances by answering questions about videos

Anonymous ECCV submission

Paper ID 5

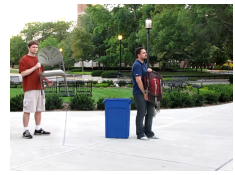
Understanding affordances in a grounded way, and using perception to acquire knowledge about affordances, requires a conceptual framework to integrate a wide range of different domains. It requires combining fields such as object recognition, human-object integration, event recognition, attention, and learning about affordances among many others. Combining these different fields is a difficult endeavour, not least because many of them operate at different levels of abstraction. For example, recognizing the functional affordances of objects likely requires tracking and understanding the relationships between multiple objects, which itself requires object detection. Beyond the raw mechanics of creating frameworks which span such a wide range of levels of abstraction and knowledge representations, such frameworks must also address multiple questions at once, otherwise we risk reinventing the wheel every time we want to take advantage of a new piece of knowledge. To this end we have been working on a new vision-language task to provide a framework to integrate knowledge from multiple domains: answering questions about images and videos.

Answering questions could serve as a unifying framework and task because it provides a flexible interface to address multiple concerns simultaneously. Chiefly, language provides a mechanism by which abstract knowledge from multiple domains, for example knowledge about events and objects, can be integrated together into a single representation. In part because of this desire to exploit priors from language, language generation from images and video has received significant attention recently. The difference between generation and question answering is important for integrating knowledge about affordances. Generation attempts to say something that is true of an image or a video, it does not produce a targeted description which must exploit knowledge about a scene.

We have created an initial approach which is able to answer questions about videos and integrates knowledge from object detectors, human-object integration, event recognition, and simple object affordances. This approach extends that of Siddharth, Barbu, and Siskind (2014) [1]. That work generates sentences which are true of a video given the grammar of a language and the semantics of words represented as HMMs. It creates a factorial HMM which jointly detects objects, tracks them, and determines if the sentence is true. Furthermore, it is able to generate sentences which are true of a video by optimally searching the space of all sentences generated by a recursive grammar.

Question answering does not merely involve generating a true statement, even if that statement is true of the objects or events referred to by the answer to the question. In Figure 1, we show two videos each depicting multiple simultaneous events. We ask a question about each video and automatically generate an


Who picked up the backpack?
The person to the right of the bin.




Who approached the bin?
The person carrying the backpack.

Fig. 1. Selected frames from two videos along with natural-language questions and automatically generated answers.

answer. Note that in both cases the answer *The person* would be correct but uninformative. You might even call it rude if a human provided this answer, because there are multiple people in the scene and clearly the questioner is asking that they be distinguished. This is not just a matter of providing an answer that is verbose, *The tall man wearing shoes* is a far more verbose answer but is equally unhelpful. What we require is discriminative sentence generation which takes into account the entire scene, the information available in the question, and the information sought out by the question.

Our approach approach searches not just for a sentence that is true but one that is also discrminiative, it is not true of other objects or events in the scene. A query is processed and a factorial HMM similar to that of Siddharth *et al.* is created to encode the information available in the question. We then use an NLP system to determine what form the answer must take. For example, is it a verb phrase, noun phase, or prepositional phrase? A new sentence generation algorithm produces a phrase of the required type, and by extension a new factorial HMM which can recognize that sentence in a video. This new algorithm contains a linear combination of two terms, one which rewards phrases that are true and one which rewards phrases that are discriminative. Two examples of the output of this algorithm are shown in Figure 1.

This new task goes beyond the standard classification and detection paradigms which dominate much of computer vision and will hopefully force a deeper understanding of scenes and videos along with integration between multiple approaches and domains. Event this early approach combines together knowledge about objects, interactions, spatial relationships, and events. We are currently extending this work to integrate knowledge about object affordances such as whether an object can be picked up and the fact that objects that are picked up must be grasped by hands. We plan to release a video question-answering corpus about objects, events, relationships, and affordances to the wider community in the near future.

References

1. Siddharth, N., Barbu, A., Siskind, J.M.: Seeing What You're Told: Sentence-Guided Activity Recognition In Video. In: CVPR. (2014)