# Research statement

Andrei Barbu

November 28, 2013

## Event recognition

We have developed an array of approaches to event recognition that ultimately combine object detection, tracking, event recognition, and sentence generation. In the base system, we develop an approach to detection-based tracking where an object detector is tuned to overdetect and detections are arranged in a lattice where links between detections in adjacent frames are scored by their motion coherence computed using optical flow. A dynamic-programming algorithm, the Viterbi algorithm, is used to find optimal paths through this lattice. The paths correspond to tracks with both high motion coherence and high-scoring detections. We call this approach the *Viterbi tracker* and present extensions for combining multiple detection sources and for producing multiple tracks per object class. Actions are recognized using hidden Markov models (HMMs) and sentences are generated using a template-based system. We develop a system which recognizes 48 actions and 33 objects and generates rich sentences containing nouns, verbs, adjectives, adverbs, adjuncts, prepositions, and determiners.

This approach is extended in a number of ways. We formulate the *event tracker* which combines tracking with event recognition by exploiting their shared structure. Recognizing the occurrence of an event involves finding the optimal state sequence through a lattice of HMM states in the same way that producing a track requires finding an optimal path through a lattice of object detections. We take the cross product of the tracker lattice and the event-recognizer lattice and simultaneously find the optimal track for the optimal event. This guides the tracker to the correct objects while at the same time recognizing the event.

Next, we formulate the *Felzenszwalb-Viterbi tracker* which simultaneously detects objects, tracks participants, and recognizes events. In principle, this is similar to the *event tracker* except that we score every possible detection in each frame rather than limiting the tracker to a small number of (over)detections returned by an object detector. In practice, the Viterbi algorithm is quadratic in the number of detections per frame. The requisite millions of detections per frame make this approach infeasible. To resolve this, we present a way of using a multidimensional generalized distance transform to produce optimal detections, tracks, and optimal HMM state sequences in time linear in the number of detections per frame. This approach allows the tracker and event recognizer to overrule the object detector and integrates all three into one generalized objective function allowing information to flow bottom-up and top down.

We then formulate the *sentence tracker* which combines tracking, event recognition, and sentence generation. This is an extension of the *event tracker* which constructs cross-product lattices from multiple tracker lattices and word lattices and constrains these cross-products to encode the co-reference and predicate-argument relations present in a sentence. In essence, this approach scores a video-sentence pair. We exploit this capability to perform 3 tasks: inference, generation, and recognition. In the first, we can guide the attention of a tracker to a particular event with a natural-language sentence despite multiple events occurring simultaneously in the field of view. In the second, we use this approach to generate sentences which describe a video. Previous approaches to sentence generation are ad-hoc, generating some sentence that describes a video. In contrast, the approach here searches the unbounded space of all possible sentences and returns the provably optimal sentence. Finally, we show how such an approach can be used to perform a novel kind of video retrieval. We search a corpus of 10 full-length Hollywood movies for clips which match a sentential query, for example *The person quickly rode the horse leftward away from the other person.* Previous approaches treated such queries as conjunctions of words and were unable to distinguish the sentence *The person rode the horse* from the sentence *The horse rode the person.* We also show how each part of speech in the query sentence impacts the results.

Finally we demonstrate *George* which performs the above in real time from live camera input.

This work was funded by the DARPA Mind's Eye program and it won both yearly evaluations of that program competing against 11 other research teams. I participated extensively in this program, attending PI meetings and one-on-one meetings with program manager. I also managed the yearly evaluations across two other universities and helped draft this and other large multi-institution proposals.

The approach taken by this work is an instantiation of a more general idea: by unifying the representations and inference algorithms of low-level perceptual problems and high-level cognition we can reason about both jointly. This opens the door to other research such as basing this approach on quadratic programming instead of the Viterbi algorithm which would allow it to integrate with state of the art segmentation approaches, obviating the need for an object detector and performing simultaneous segmentation, tracking, and event recognition. Extending this work to 3D, which is required to recognize many events, is also of interest and will likely require integrating reasoning about a projection model and about 3D spatial relations of objects. One of the results of such work would be more reliable 3D object detectors which are lacking at present. This work has so far explored events which are described by a sentence, but can in principle be extended to consider events described by a paragraph or story which would enable tackling a number of new problems. One such problem is a new approach to planning: planning by imagination by reasoning about filling gaps in videos and stories using language generation.

Collaborators: Sven Dickinson (Toronto) and Song Wang (South Carolina).

Publications: [J2], [J3], [J5], [C1], [C2], [C3], [C6], and [T1]

## Compositional representation of events in the human brain

How does the human brain represent simple compositions of objects, actors, and actions? The fact that this representation is compositional is taken for granted by many in the cognitive-science and artificial-intelligence communities. For example, in computer vision, representations for nouns, such as those used for object detection, are independent of representations for verbs, such as those used for event recognition.

To explore if humans employ compositional representations, we had subjects view action-sequence videos during neuroimaging (fMRI) sessions and decoded the resulting activation patterns to label the videos. In other words, we took, as input, the fMRI activation patterns and recognized the activity in the video. We first decoded labels for the videos corresponding to one of six verbs: *carry*, *dig*, *hold*, *pick up*, *put down*, and *walk*. This decoding is reliable with a mean accuracy of 65%, where chance is 16.66%. This is the first experiment which decodes labels corresponding to verbs; earlier work had focused on nouns which are represented in different parts of the brain, processed by different pathways, and require different stimuli. This novel ability was then used to recover sentential descriptions of videos. We showed subjects videos which depict three verbs (*carry*, *fold*, and *leave*), each performed with three objects (*chair*, *shirt*, and *tortilla*), each performed by four different human actors, and each performed on either side of the field of view. We then recovered sentential descriptions of the form 'the *actor verb* the *object direction/location*' by separately recovering each lexical component. On this is 1-out-of-72 classification task, we decode the correct sentence with 13% accuracy (chance 1.3%). This is the first experiment which has demonstrated an ability to decode a complex concept with multiple components. If the representation of complex concepts is compositional, at least at the level at which current neuroimaging techniques allow us to investigate, we expect that the performance of a classifier trained separately on nouns and verbs should match the performance of a classifier trained jointly on both. In other words, if the representation is compositional we expect that decoding the noun will provide little to no information about the verb, and vice versa, given that our corpus is designed to allow each action to be performed with each object. Indeed, this is what we find. The performance of a classifier trained jointly on nouns and verbs (50% accuracy) essentially equals the performance of classifiers trained separately on the two (48% accuracy), on a task where chance is 16%. Moreover the brain regions from which the classifier derives most of its performance are largely disjoint between nouns and verbs, and the joint noun-verb classifier uses essentially the union of these two regions. Furthermore, this same analysis can be performed on all pairs and triples of actor, verb, object, direction, and location yielding essentially the same results. This is the first evidence for the compositional nature of representations in the human brain.

This research brings up the possibility of systematically exploring the neural basis of different linguistic theories. In addition, earlier work has shown that nouns have internal structure and their representation in the brain seems to follow radial categories. This work opens up the possibility of exploring the internal structure of verb representations. It would also be interesting to replicate this work with other animals and investigate the nature of neural representations and their relationship to language in non-humans.

Collaborators: Jason Corso (SUNY Buffalo), Steven Hanson (Rutgers), Barak Pearlmutter (NUI Maynooth), Tom Talavage (Purdue), and Ronnie Wilbur (Purdue).

Publications: [J1]

## Learning physically-instantiated game play through visual observation

Children learn to play games by visual observation without generally being told the rules of the game. Even as adults, we rarely read the rulebook of a game and begin to play. We generally either play a practice game or watch someone play the game first and then understand the rules. Moreover, outside the constrained world of board games, social interactions don't come with a rulebook, or even a fixed set of rules, and we must learn the rules governing how we interact with each other largely by observation.

To understand how a robotic system can perform these learning tasks, we have developed an approach to learning to play board games. This provides a microcosm to explore concepts such as teaching, competition, and strategy. Unlike previous work, we focus on acquiring the rules necessary to play legally, not necessarily well. Two robots are provided with the rules of a board game expressed in natural language. They use these rules to play a physical board game while a third robot watches. This third robot, the learner, does not know the rules of the game but has background knowledge about the world that would be available to any child, such as basic spatial relations. The learner combines positive evidence from visual observation, as well as inferred negative evidence, with the background knowledge, to learn the rules of the game. Negative evidence can be inferred, for example, by the fact that every time you see that a game has not ended you know no one has won. The rules are learned from a small number of examples, typically 3 to 10. The learner then uses these learned rules to engage in physical play.

In the future, this platform can be used to explore more general notions which transcend particular games, such as the concepts of attack, defend, and area control. These concepts need not be expressed in board games, and I would like to investigate how knowledge transfer can happen between very different kinds of games such as chess and soccer. The ability to learn to play games also enables integrating low-level perception with high-level cognition by jointly reasoning about how both can affect the possible legal game rules in order to automatically detect the board, as well as its connectivity, and the pieces, as well as their features such as their ability to stack. In addition, I would like to explore how robots and humans can teach one another.

Publications: [C8]

## Seeing, describing, and manipulating part-based 3D structures

Humans can recognize complex structures, such as engines, which are composed of multiple parts. We can discuss these structures, describe them to others, and manipulate them. Children perform similar tasks with assembly toys such as Lego. They recognize structures, discuss different aspects of those structures, modify structures, and build structures. All of these tasks require combining language, 3D vision, knowledge about the physics of the world, and robotics in a challenging domain. 3D part-based structures are extremely difficult to recognize. From any view, most of the structure is occluded. Moreover, many parts tend to have the same visual characteristics making recognizing the presence or absence of a part a difficult task.

We develop a unified approach which can see, describe, and manipulate part-based 3D structures. It uses knowledge about the possible 3D parts and about the physics required to assemble stable structures to recover the part-by-part composition of a structure. This allows one to infer knowledge about the structure that cannot be seen in the image, for example the existence of a crucial beam which is occluded but is required to form a stable structure. To achieve this, a single graphical model combines knowledge of the 3D parts, physical knowledge of stable structures, occlusion, and language. This graphical model is expressed in a probabilistic programming language. Methods to perform exact inference for this model are developed despite its immense size. This model naturally combines multiple sources of evidence, like language and vision, and allows for an interactive system which determines whether it is confident in its current estimated structure and what future action should be taken to raise that confidence. For example, the system can image a part of the structure that is occluded, and known to be undetermined, by moving its camera or disassembling part of the structure. Multiple sources of such visual evidence can be combined along with linguistic evidence in the form of natural-language sentences which describe part of the structure. This framework can be used to build structures from linguistic input, describe visually observed structures in language, explore structures to understand more about them, and disassemble structures to understand their internals. We demonstrate this approach using a robotic arm and Lincoln Logs, a children's assembly toy.

This work has so far explored a number of questions:

1. How do you know what is occluded, which requires knowing what the structure is, in order to determine the structure?
2. How do you measure confidence in an estimated structure?
3. What is the optimal sequence of actions to increase your confidence?
4. How do you describe a part-based structure?
5. How do you combine incomplete sentential descriptions of structures with multiple images from different views of the structure in different states of assembly in order to estimate the structure?

It also opens the way for addressing other interesting questions such as:

1. How can knowledge about physics be acquired?
2. How can the knowledge about the affordances of 3D objects be acquired?
3. How does low-level kinematics interact with the high-level control required to manipulate the parts?
4. How does this scale to other domains which have similar structure instantiated in different ways, such as electronic circuits or mechanical devices?
5. How can knowledge be transferred across domains, such as learning to build or recognize a structure using one set of parts and then constructing the same structure out of a different set of parts?

Publications: [J4] and [C7]