Video Activity Recognition in the Real World

Anthony Hoogs and **A. G. Amitha Perera**
Kitware, Inc.
28 Corporate Drive, Clifton Park, NY 12065
anthony.hoogs@kitware.com amitha.perera@kitware.com

Abstract

With recent advances in motion detection and tracking in video, more efforts are being directed at higher-level video analysis such as recognizing actions, events and activities. One of the more challenging problems is recognizing activities that involve multiple people and/or vehicles, whose relationships change over time, when motion detection and tracking are unreliable, as commonly occurs in busy scenes. We describe an approach to this problem based on Dynamic Bayesian Networks, and show how DBNs can be extended to compensate for track failures. We also show that defining DBNs with semantic concepts improves robustness vs. direct observables, and discuss implications and ideas for incorporating semantic, symbolic knowledge into the perceptual domain of activity recognition.

Introduction

As motion detection and tracking in video has matured over the past ten years, research in various forms of motion and track analysis has significantly increased. The most popular topics include human action recognition, which focuses on single-person, short-duration actions such as walking, running, jumping, bending and drinking (Laptev et al. 2007); motion pattern learning and anomaly detection, which attempts to detect when a (single) moving object is not behaving normally (Makris and Ellis 2005; Swears, Hoogs, and Perera 2008); activity recognition, which models and detects specific, dynamic interactions between movers (Xiang and Gong 2005; Hongeng, Nevatia, and Bremond 2004). Research in the AI community (Liao, Fox, and Kautz 2005) has been focused on track-level analysis of single object activity.

In this paper we are concerned with activity recognition, and in particular the modeling and recognition in video of *complex* activities that involve multiple agents and multiple temporal states. One example of such an activity, aircraft refueling, is shown in Figure 1. This activity involves two movers (the fuel truck and the refueler) and the aircraft itself, and is modeled with 17 temporal states that capture the sequence of observables corresponding to the various stages of a refueling operation. Other complex activities in the aircraft domain include baggage loading/unloading, food servicing, passengers embarking/disembarking, de-icing, and aircraft parking. Activities commonly examined in other domains include stealing, fighting, exchanging packages, and covert surveillance.

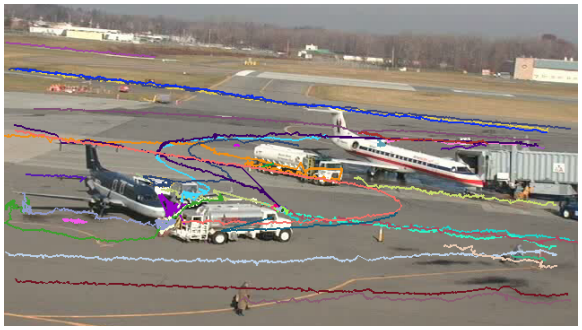

Figure 1: An example of a multi-agent, multi-state activity: refueling of an aircraft. Computed tracks from the time interval containing the activity are overlaid on a frame of the video. Each track has a unique color. Two moving objects are involved in the activity, and they generate 7 tracks. The remaining tracks are from clutter activities.

Of the various forms of motion and track analysis, complex activity recognition has perhaps the most relevance to AI, as it can be viewed as a form of plan recognition and typically requires a higher-level knowledge representation in some form. However, video activity recognition approaches tend to be driven by the observables, rather than models of underlying intent or purpose. This occurs because of the difficulty in dealing with video observables; there are various types of tracking errors, some of which are evident in Figure 1, that greatly increase the difficult of activity recognition based on tracking. Extracted tracks are often fragmented, meaning that the trajectory of one object is broken into multiple tracks. Tracks are frequently missing as objects become occluded or wander in and out of the camera field of view. When objects are close together, one track may incorporate multiple objects. Perhaps the most difficult error for activity recognition is track switching, where a single track jumps from one object to another without breaking.

Researchers have recently begun to address these problems. One solution is to avoid tracking entirely and rely only on moving object detections, which are relatively reliable (Gong and Xiang 2003; Xiang and Gong 2005). However, they introduce spatial dependence and model localization to compensate for the lack of tracking, and hence construct models that are specific to each scene. Other methods simply assume that tracks of activity agents are complete through the duration of the activity, and degrade rapidly when this assumption fails (Hongeng, Nevatia, and Bremond 2004; Intille and Bobick 1999).

Figure 2: Another example of a refueling activity with a different viewpoint and scene configuration. The lines show manually specified object tracks.

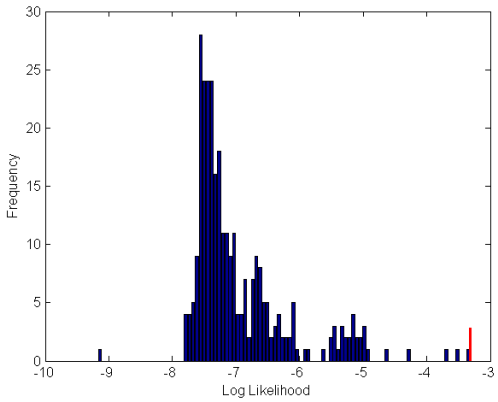

Figure 3: Histogram of log-likelihood scores calculated on observation sequences derived from a testing dataset. 303 different observation sequences were generated by assigning different actor roles to manually specified object tracks. The 303 role assignments include 300 random ones, and 3 hand-picked ones that were meant to be confusable. The red bar corresponds to the correct role assignment and was found to be well separated from the others.

In our recent work, we have studied the effects of tracking errors on complex activity recognition, and proposed a solution that jointly solves activity recognition and tracking together (Chan et al. 2006a; 2006b). We adopt a more AI-centric approach than some (Xiang and Gong 2005) by modeling the underlying semantics of the activity using the Cyc ontology (Lenat and Guha 1990), and transform direct observables such as object position and velocity into low-level semantic primitives before incorporating them into the model. Most significantly, despite its symbolic nature our method compensates for typical observable errors by judiciously exploring the combinatorial hypothesis space of activity type, assigning track segments to activity roles, temporal interval (start and end times), and state durations.

Activity recognition under such real-world conditions is yet another manifestation of the classic signal-to-symbol problem, also known as the semantic gap. Of particular significance here is that the problem is temporal in nature; multiple agents are interacting; and the observables are error-prone and inherently under-specified by photographic projection. We expect that our methods could apply to generic plan recognition problems, as well as activity recognition beyond the video domain. Our approach is summarized in the next section, followed by a discussion of the broader implications for AI.

Activity Recognition with Tracking Errors

In our work (Chan et al. 2006a; 2006b), an activity model consists of a fixed set of actors and a dynamical model expressing how the actors interact over time. Each actor has an associated track representing its motion over time. We use a Dynamic Bayesian Network (DBN) to represent the dynamical model, and build the DBN as a progression of intermediate semantic states obtained from the Cyc ontology (Lenat and Guha 1990). As an example, one of the intermediate states for the "refueling" concept is "TransportWithMotorizedLandVehicle", describing the movement of the fuel truck approaching the airplane.

The observable nodes in the DBN are derived from a set of Cyc concepts describing spatial and temporal concepts on single objects and object pairs; examples include "close to", "moving", "contained in", and "disappear near". Our set of predicates is analogous to those used by others for activity recognition (Hongeng, Nevatia, and Bremond 2004; Intille and Bobick 1999). Using these relations instead of the raw track measurements yields a great degree of robustness against changes in viewpoint and scene configuration (Chan et al. 2004), allowing the same model to be applied to different scenes. For example, a model derived from the data shown in Figure 2 can be applied to that shown in Figure 1. In contrast, approaches that use the raw track or object location measurements tend to learn models that encode specific locations for certain actors, and hence are difficult apply to beyond that specific configuration and viewpoint.

To recognize the activity, we first use a general video tracking algorithm to track all objects in the scene. We then assign a track to each actor defined in the model, resulting in an *observation sequence*: a sequence of observations for each actor in the model. If a particular observation sequence is explained by the DBN with high enough likelihood, we say that the corresponding tracks executed the modeled activity. To recognize the activity, we evaluate the model over all possible observation sequences in a brute force manner. Figure 3 shows the histogram of likelihoods for a single model over a set of different (and incorrect) track-actor assignments. Here, we used manually specified tracks to ensure that the tracks were complete. The dataset used to create the activity model is significantly different in view point and scene configuration from the dataset used to generate the histogram. Even so, the likelihood of the model for the true assignment (red line) is well separated from the rest, indicating that the model can indeed reliably recognize the activity, and that the model generalizes well to different scenes.

Thus far, we have assumed that the trajectories for all the

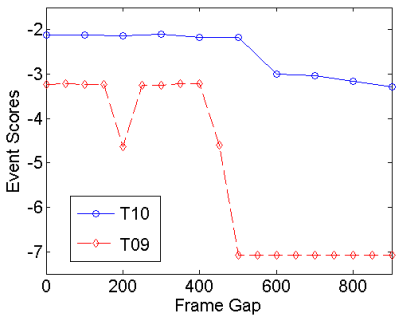

Figure 4: Results of recognizing the refueling activity on manual tracks, with increasing levels of fragmentation. The horizontal axis is the length of track gaps, and the vertical axis is the average log-likelihood. T10 is the sequence used for training. Reasonable recognition performance is maintained on sequence T09 over track gaps up to 400 frames.

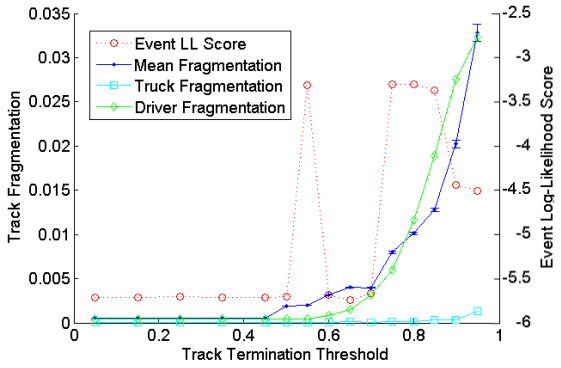

Figure 5: Fragmentation and activity log-likelihood score L as functions of termination threshold τ. Fragmentation for the driver and all objects have very similar dependence on τ; the truck has a similar profile with much smaller scale. L increases dramatically when the driver fragmentation increases slightly at $\tau = 0.75$.

actors are complete; an assumption that is also widely made in the computer vision activity recognition literature. However, as we stated in the introduction, this is hardly ever true with real data. Video object trackers can and do fail for any number of reasons, including noise, occlusions, and unexpected multi-object interaction. The most common failure type is early termination: a track is terminated even though the corresponding object is still present. However, if the tracker is aggressive, it may try to track through temporary occlusions, and will sometimes latch onto a different object than was occluded. This results in a track-switching error, where part of the track corresponds from one object and the rest to a different object.

Track switching errors are often disastrous for high-level reasoning, because the models generally assume that each computed track corresponds to a single object. In contrast, early termination errors are far more benign, because the the resulting track fragments can be subsequently linked together to form longer trajectories, albeit with gaps (Perera et al. 2006). Figure 4 shows that the activity can be recognized even with significant gaps in the trajectory.

The tendency of a tracker to terminate early can be characterized by a termination threshold τ, with $\tau = 1$ yielding a very conservative tracker and $\tau = 0$ a very aggressive one. A conservative tracker tends to stop tracking whenever there is the slightest chance of tracking error, and thus produces very short tracks. When these are linked, the result is highly fragmented trajectories. An aggressive tracker, on the other hand, produces longer tracks and less fragmentation, but is more likely to make switching errors. The video tracking community generally considers longer tracks to be "better" tracking; however, Figure 5 shows that activity recognition favors shorter tracks, if there is a linking step available to join the fragments into longer trajectories. This is because activity recognition is tolerant of gaps in the data, but is sensitive to switching errors.

Typically, track fragments are linked together based on kinematic and appearance constraints, and even with excellent track linking, the linked trajectory will still be incomplete. We have recently begun exploring the notion of using the activity model to aid the linking, by linking together tracks that cause the linked trajectory to better fit the model. For this to work, the model must be able to recognize partial activities: the likelihood scores from partial but correctly assigned tracks must be better than incorrectly assigned tracks. Our initial results indicate that this is indeed true. For example, in one experiment, the trajectory of a key actor in a refuelling activity was divided into three fragments even after track linking. We ran the activity model for each fragment, as if the fragment described the complete trajectory of the actor. This resulted in likelihood scores of -5.2, -4.7, and -4.2. Comparing to the scores in Figure 3, these numbers indicate partial recognition.

Our initial attempt (Chan et al. 2006a) tried to incorporate the DBN evaluation into our mutli-object tracking framework (Perera et al. 2006) because of the latter's efficiency. However, the resulting framework has significant drawbacks such as limiting the track linking solution to one actor track at a time. An efficient mechanism to explore the solution space is required, as described in the next section. We are currently exploring an adaptation of multiple hypothesis tracking to track fragments, and trying to leverage the temporal nature of the problem to our advantage.

Video Activity Recognition and AI

Much of the video analysis research conducted in the computer vision community has little or no consideration of traditional AI (as distinguished from machine learning, which is highly integrated with vision). As video research matures and moves towards higher levels of semantic analysis, however, there is greater opportunity for vision and AI to benefit from each other. In particular, video activity recognition usually requires or implies understanding of scene context, actor intent and even common-sense knowledge. Increasing the innate "intelligence" of activity recognition should lead to significant performance improvements and general-

ization.

One of the primary barriers in using higher levels of knowledge in vision is the challenge of high-dimensional model representation, efficient matching and contextual reasoning in the presence of highly uncertain and missing observations. Purely symbolic methods do not work well, and indeed many would credit recent advances in vision to the extensive use of probabilistic models.

In our activity recognition work, we follow recent trends with the use of DBNs to represent activities. DBNs have received significant attention lately within the video and machine learning communities because of their combination of representational power and handling of observation noise and gaps. However, our approach illustrates that DBNs alone do not solve the video activity recognition problem, because they do not provide a method for exploring much of the search problem underlying activity recognition, or a direct means for large-scale contextual reasoning.

The search space is huge, and is roughly the product of the number of activity models; actors per model; track fragments per actor; possible temporal intervals for the activities; and tracks in the scene. The state-of-the-art in the vision literature is to prune the data until most of those dimensions reduces to 1, and perform a brute force search over the remaining space, which is typically over the actor-track assignment space.

In the primary contribution of our work, we attempted to address the issue of incomplete trajectories. This needs to be solved for activity recognition to work in the real world, because the tracks will always be fragmented. We have shown that activity recognition can be successfully performed even with fragmented tracks, but our initial attempt at an efficient solution placed too many constraints on the model evaluation.

A second consideration in our work is the use of prior structured knowledge with uncertain observables. Many activity recognition methods, particularly those addressing complex activity recognition, implicitly define an ontology specific to the problem at hand without acknowledging this fact. Only a few researchers have explicitly defined ontologies for video analysis (Nevatia, Hobbs, and Bolles 2004; Maillot, Thonnat, and Boucher 2003). We desired to use an existing large-scale ontology to enable large-scale reasoning, model generalization across domains and intuitive model construction—while handling noisy, perceptual observables.

To avoid creating our own toy ontology, we conceptually grounded our DBN representation within the Cyc ontology (Lenat and Guha 1990), as described above.

By defining the DBN with semantic predicates as observables, we enabled small-scale probabilistic reasoning on semantic concepts at the predicate level, and also at the state concept level. What we did not do was to automatically associate these two semantic levels. Within Cyc, there exist paths between each state concept and the predicates that define its observables. Since predicates are computable from observables, it may be possible to automatically determine which predicates should be observed for an arbitrary, high-level Cyc concept given sufficient context (e.g. other concepts in sequence). That would enable the generation of computable activity models by the simple selection of a sequence of intuitive, human-level Cyc concepts.

We also would like to examine the scalability of probabilistic reasoning within a large ontology, linked to perceptual observables such as tracks. We hope to use commonsense reasoning during recognition to reduce false alarms and incorporate local, domain-specific context.

References

Chan, M.; Hoogs, A.; Schmiederer, J.; and Petersen, M. 2004. Detecting rare events in video using semantic primitives with HMM. In *Proc. ICPR*, volume 4, 150–154.

Chan, M. T.; Hoogs, A.; Bhotika, R.; Perera, A. G. A.; Schmiederer, J.; and Doretto, G. 2006a. Joint recognition of complex events and track matching. In *Proc. CVPR*.

Chan, M.; Hoogs, A.; Sun, Z.; Schmiederer, J.; Bhotika, R.; and Doretto, G. 2006b. Event recognition with fragmented object tracks. In *Proc. ICPR*.

Gong, S., and Xiang, T. 2003. Recognition of group activities using dynamic probabilistic networks. In *Proc. ICCV*, 742–749.

Hongeng, S.; Nevatia, R.; and Bremond, F. 2004. Video-based event recognition: activity representation and probabilistic recognition methods. *CVIU* 96(2):129–162.

Intille, S., and Bobick, A. 1999. A framework for recognizing multi-agent action from visual evidence. In *Proc. Natl. Conf. on AI*, 518–525.

Laptev, I.; Caputo, B.; Schüldt, C.; and Lindeberg, T. 2007. Local velocity-adapted motion events for spatio-temporal recognition. *CVIU* 108(3):207–229.

Lenat, D., and Guha, R. 1990. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Reading, MA: Addison-Wesley.

Liao, L.; Fox, D.; and Kautz, H. 2005. Location-based activity recognition. In *Advances in Neural Information Processing Systems (NIPS)*.

Maillot, N.; Thonnat, M.; and Boucher, A. 2003. Towards ontology-based cognitive vision. In *Proc. Intl. Conf. on Computer Vision Systems*.

Makris, D., and Ellis, T. 2005. Learning semantic scene models from observing activity in visual surveillance. *IEEE T. SMC-B* 35(3).

Nevatia, R.; Hobbs, J.; and Bolles, B. 2004. An ontology for video event recognition. In *Proc. IEEE W. on Event Detection and Recognition*.

Perera, A.; Srinivas, C.; Hoogs, A.; Brooksby, G.; and Hu, W. 2006. Multi-object tracking through simultaneous long occlusions and split-merge conditions. In *Proc. CVPR*.

Swears, E.; Hoogs, A.; and Perera, A. G. A. 2008. Learning motion patterns in surveillance video using hmm clustering. In *WMVC*.

Xiang, T., and Gong, S. 2005. Video behaviour profiling and abnormality detection without manual labelling. In *Proc. ICCV*, volume 2, 1238–1245.