An Ontology based Approach for Activity Recognition from Video

Umut Akdemir
University of Maryland
College Park, MD
uakdemir@cs.umd.edu

Pavan Turaga
University of Maryland
College Park, MD
pturaga@umiacs.umd.edu

Rama Chellappa
University of Maryland
College Park, MD
rama@umiacs.umd.edu

ABSTRACT

Representation and recognition of human activities is an important problem for video surveillance and security applications. Considering the wide variety of settings in which surveillance systems are being deployed, it is necessary to create a common knowledge-base or ontology of human activities. Most current attempts at ontology design in computer vision for human activities have been empirical in nature. In this paper, we present a more systematic approach to address the problem of designing ontologies for visual activity recognition. We draw on general ontology design principles and adapt them to the specific domain of human activity ontologies. Then, we discuss qualitative evaluation principles and provide several examples from existing ontologies and how they can be improved upon. Finally, we demonstrate quantitatively in terms of recognition performance, the efficacy and validity of our approach for bank and airport tarmac surveillance domains.

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: Video Analysis—*Human Activity Ontologies*

General Terms

Algorithms, Performance, Standardization

Keywords

Activity Ontologies, Visual Surveillance

1. INTRODUCTION

Designing algorithms that can recognize human activities in video sequences has been an active field of research during the past ten years. With the proliferation of visual surveillance systems in a wide variety of domains such as banks, airports, convenience stores etc, it is necessary to create a general representation framework for modeling activities. Examples of such knowledge-bases, or more formally, ontologies have been in existence in other fields of AI such as the semantic web, image and video annotation and computational genomics. Designing ontologies for human activities has only recently been gaining interest in the computer vision community. The advantages of such a centralized representation are easily seen. They standardize activity definitions, allow for easy portability to specific deployments, enable interoperability of different systems and allow easy replication and comparison of system performance. In the vision community, ontologies have mostly been tuned for specific scenarios. We build upon these efforts and discuss design and evaluation principles to facilitate the process of designing a generalizable ontology for human activities.

Related Work Statistical pattern recognition approaches such as HMMs have long been the mainstay in computer vision for modeling and recognizing atomic actions. Statistical approaches require a training phase where models are learnt from training exemplars. But, a rich training set which encompasses all possible manifestations of an activity may not exist. Further, complex multi-agent activities require more sophisticated syntactic and structural approaches such as dynamic belief networks [8], logic networks [7] and context free grammars [9]. Structural approaches usually rely on hand-crafted models from experts or analysts, hence they are intuitive and can be related to the semantic structure of the activity. Since, these approaches rely on a domain expert to provide the activity semantics, it is useful to create a standardized knowledge-base from which to draw upon. Recent efforts have focused at creating knowledge-based semantic descriptions or 'ontologies' for activities. Examples include the works of [1] who use ontologies for analyzing social interaction in nursing homes, [4] who used ontologies for classification of meeting videos and [2] who use ontologies to recognize activities in a bank monitoring setting. As a result of the Video Event Challenge Workshops held in 2003 [1], ontologies have been defined for six domains of video surveillance - 1) Perimeter and Internal Security, 2) Railroad Crossing Surveillance, 3) Visual Bank Monitoring, 4) Visual Metro Monitoring, 5) Store Security, 6) Airport-Tarmac Security. The workshop led to the development of two formal languages – Video Event Representation Language (VERL) [5], which provides an ontological representation of complex events in terms of simpler sub-events, and, Video Event Markup Language (VEML) which is used to annotate VERL events in videos.

In most practical deployments, activity definitions are constructed in an empirical or ad-hoc manner. Though empirical constructs are fast to design and even work very well in most cases, they are limited in their utility to the specific deployment for which they have been designed. In the context of human activities, while it is easy to see the advantages of a task-independent knowledge-base, we will show that it is also necessary to make task-dependent simplifications.

[1]Event Ontology Workshop, 2003 http://www.ai.sri.com/∼burns/EventOntology

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

Table 1: Cruise Parking Lot

```
PROCESS(cruise-parking-lot(vehicle v, parking-lot lot),
Sequence(enter(v, lot),
    set-to-zero(i),
    Repeat-Until(
        AND(move-in-circuit(v), inside(v, lot), increment(i)),
        equal(i, n)),
    exit(v, lot)))
```

Organization of the paper: In section 2, we discuss ontological design issues for human activities. Then in section 3, we discuss qualitative evaluation criteria for ontologies and provide examples from existing ontologies. Then, in section 4, we apply these principles on two domains and show how they improve the recognition performance. Finally, in section 5 we provide concluding remarks.

2. ONTOLOGY FOR ACTIVITIES

An ontology for human activities describes entities, environment, interaction between them and the sequence of events that is semantically identified with an activity. It specifies how an activity can be constructed using lower-level primitive events by identifying the role played by each entity in the sequence of events. Human activities are characterized by complex spatio-temporal interactions of primitives and contextual entities. Consider the activity of a car cruising in a parking lot whose definition is given in table 1. This definition illustrates the importance of encoding the temporal and spatial constraints. Without special regard to spatio-temporal constraints, we might say that it is composed of the primitives: 'car_enter', 'car_move_in_circuit' and 'car_exit'. But, a normal car-parking activity can also be described using the same set of primitives as car-cruising. The difference that sets apart the two activities lies in the temporal span of the 'car_move_in_circuit' primitive.

3. EVALUATING ONTOLOGIES

Thomas Gruber [3] made one of the earliest systematic attempts to propose guidelines for the design and development of ontologies intended for knowledge sharing and interoperability. Five important criteria for ontology design - clarity, coherence, extendibility, minimal encoding bias and minimal ontological commitment were proposed.

1. Clarity: An ontology should convey the meaning of all conceptualizations unambiguously.
2. Coherence: An ontology should be coherent and allow for meaningful inferences to be drawn that are consistent with definitions and axioms.
3. Extendibility: The design of the ontology should take into account future extensions without a need for revising definitions.
4. Minimal Encoding Bias: An ontology with minimal encoding bias has conceptualizations that are symbol-independent.
5. Minimal Ontological Commitment: An ontology makes as few assumptions as possible about the domain being modeled.

In the following discussion we examine some examples from the Event Ontology workshop output to show how to resolve ambiguities according to the design principles outlined above.

Table 2: Tail-gate definition in Perimeter and Internal Security

```
SINGLE-THREAD(tailgate(ent x, ent y, facility f)
    AND(portal-of(entrance, f)),
    Sequence(AND(approach(x, y)behind(x,y)),
    tail-behind(x, y),
    get-access(y, entrance),
    enter(y,facility),
    NOT(get-access(x, entrance)),
    enter(x, facility)))
```

Table 3: Shoplifting in Store Security Ontology

```
SINGLE-THREAD(shop-lift (person x, employee y, ent o),
AND(counter(area),
    merchandise(o),
    Close(area),
    NOT present(y,area),
    Sequence(move(x,area),
    Open(area),
    Pick-up(x,o));
```

Example 1: Clarity in Temporal Relations: In the Perimeter and Internal Security Ontology, tailgating is defined as in table 2. This activity is characterized mainly by sequentiality of primitives. In the current example, the 'tail-behind' activity occurs before the 'enter' primitive for entity x, and 'enter' for x occurs between the time y gains access to the entrance and the time the entrance to the facility closes. For unambiguous representation, we need definitions of 'before' and 'between' in temporal relations for example as presented in [6].

Example 2: Clarity of Negation in Time: In this example, we examine the 'shop-lifting' activity definition (table 3) which has two types of prefixes – time-bound and time-independent. Prefixes such as 'present', 'pick-up' are prefixes with specific temporal spans. For example, 'pick-up' is true only for the duration during which the pick-up action occurs. On the other hand, prefixes such as 'counter', which refers to store-counter, is an assertion that is true independent of time. This causes an ambiguity when temporal negation is expressed.

According to this definition, the shop-lifting activity involves a person x picking up merchandise o from the counter when the employee y is not present at the counter. The employee y not being present in the area is only true for a finite temporal span. It is not a truth that is independent of time. Hence, for this negation to be meaningful there should also be a corresponding time interval associated with it. For disambiguation, the time-dependent 'not' should be defined. We can define it as given in table 4. According to this definition, the prefix is not true within any sub-interval of the given time interval. By separating time-dependent negation from negations expressing falseness of a concept, we resolve the ambiguity caused by negations that are only true for a specified period. Similar arguments can also be made for the previously defined tailgating activity (table 2).

Consider now a variation of the above example and try to comprehend the meaning of time-independent NOT of a NOT IN INTERVAL as defined in table 5. We see that it corresponds to negation of the prefix not being true in any sub-interval of the given time interval. This implies that the prefix is true in at least one sub-interval of the time interval. This understanding prevents the ambiguity that can arise from negation.

Table 4: Definition of Temporal Negation

```
NOT-IN-INTERVAL(prefix (entity list ),time interval )
```

Table 5: Negation of Temporal Negation

NOT(NOT-IN-INTERVAL(prefix (entity list),time interval))

Table 6: Suspicious Load in Perimeter Security

SINGLE-THREAD(suspicious-load(vehicle v, person p, ent obj, facility fac), AND(zone(loading-area), near(loading-area, facility), portable(obj), Sequence(approach(v, fac), AND(stop(v), near (v, fac), NOT(inside(v, loading-area))), AND(approach(p, v), carry(p, obj)), AND(stop(p), near (p, v)), cause(p, open(portal-of(v))), enter (obj, v), cause(p, close(portal-of(v))), leave(v, facility))))

Example 3: Minimal Ontological Commitment: Let us consider an example from Perimeter Security ontology for a suspicious load, given in table 6. Though this example maintains clarity, further inspection shows that minimal ontological commitment is not preserved. According to this definition, for a suspicious load, the vehicle's portal has to be opened in order to load the object. This does not encompass other possible scenarios. For example, the suspicious load can be placed onto the trailer of a truck which is open from the top, hence not using any portal, or it can even be an explosive that is placed under the body of the vehicle. Moreover, it is not necessary for the vehicle to stop. For instance, somebody inside the vehicle could grab a bag from a suspicious pedestrian through the window. Hence, minimal ontology should only include the object being on the vehicle's exterior, and then being transferred to the vehicle's interior while it is in an undesignated zone.

Similarly, in the bank ontology, we observe that there are many safe-attack (an attack on the bank safe) definitions with only minor differences. In the single-threaded definitions – involving a single robber (for example, see table 8) – deviations in the path of the robber or the opening or closing of the bank entrance are used to define different activities. Moreover, there are several multi-threaded versions involving two robbers with minor differences (for example, see table 9). If we examine these variations, they are characterized by at least one of two common occurrences: 1) Either someone is hurt, or 2) there is unauthorized access to the bank safe. These are enough to minimally formulate the suspicious activities in a bank. Detection of suspicious activities can be achieved by detecting at least one of the occurrences. A simplified ontology based on this observation is given in table 7.

3.1 Granularity of Ontologies

In this section we will discuss granularity issues in ontology design and the role of context in determining the ontological complexity.

Bank Surveillance: In the bank surveillance scenario, there may be many different variations of the same activity, for instance a robbery. Distinct definitions of robberies involving single or multiple robbers, deviations in the path of the robber, state of contextual objects such as the bank entrance etc, is not a feasible solution. As discussed in the previous section, the occurrence one of two events – a) someone getting hurt, or b) an unauthorized access to the safe – suffices to distinguish the suspicious activities from the normal activities. Hence, for this domain ontological definitions at a very fine level of granularity is not required.

Table 7: Attack in Bank

safe attack: usage: safe attack(mo1,z1) physical objects: ((mo1:mobile object),(z1:zone)) components: ((c1:approach(mo1,z1)) (c2:inside zone(mo1,z1))) (c3:leave(mo1,z1))) (c4:NOT(employee((mo1)))) temporal constraints: (sequence(c1,c2,c3))

Airport Tarmac Surveillance: The airport surveillance domain is characterized by activities that demonstrate a high degree of synchronization and structure among several agents. For example, the arrival of an aircraft at the terminal is preceded by a sequence of ground-crew actions which is highly structured. There is not much room for deviations from this strict procedure. Thus, a finer level of granularity is warranted to characterize activities in this domain, than for instance, the bank domain. In this context, the ontology output from the challenge workshop is minimal – the definitions encode only the necessary and sufficient information.

4. EXPERIMENTS

Bank Dataset: The bank dataset [10] consists of six video segments, four of which contain different instances of bank robberies and the other two contain instances of normal activities in the bank. The bank scenario consists of a single camera with a relatively static background. We used background subtraction to identify moving objects in the scene. Contextual entities such as entrance to the bank, entrance to the safe, management office, service counter etc were manually marked. The moving objects are tracked using motion and color based appearance matching. Once the low-level primitives are detected, we used a simple finite-state machine to represent the activity constraints given by the definitions.

In the first experiment, we used the definition given in Table 7 to recognize bank attacks. All the 4 variations of the bank attack were recognized by this definition. Results on two videos are shown figures 1 - 2. For comparison, we used two other definitions from the Event Ontology workshop – a single-threaded ontology (STO) and a multi-threaded ontology (MTO) whose definitions are given in tables 8 and 9. All the available video segments were classified into one of two classes - a) Abnormal and b) Normal, where abnormal corresponds to an attack on the bank safe. We obtained correct classification of all available instances of normal and abnormal activities using the proposed minimized ontology. The STO correctly classifies three of the four attack scenarios, but fails on one involving two robbers and entry into the management office. The MTO definition fails to correctly recognize the instances involving not only a single robber but also the ones involving two robbers, since it is not minimal.

Table 8: Single thread Event: Attack in Bank

composite event Safe attack 1-person-back-counter: physical objects: ((p : Person), (z1: Back Counter), (z2: Safe)) components: (c1: primitive event Changes-zone (p, z1, z2))

TSA Airport Surveillance dataset: In the TSA data, several activities such as arrival and departure of aircraft, embarkation and disembarkation of passengers, etc are observed. We drew upon the activity definitions from the Event Ontology workshop for this domain. As noted in section 3.1, the ontology was found to be minimal, hence we did not try to minimize it further.

Figure 1: Bank Robbery Variation 1. (a) Person enters the bank, (b) Robber is identified to be an outsider. Robber is entering the bank safe, (c) Robber makes an exit.

Figure 2: Bank Robbery Variation 2. (a) Person enters the bank, (b) Robber is identified to be an outsider. Robber is entering the bank safe, (c) A customer escapes, (d) Robber makes an exit.

Table 9: Multi Thread Event: Attack in Bank

composite event Safe attack 2-persons-inside-safe-entrance
physical objects:
((p1: Person), (p2: Person), (z1: Safe Entrance))
components:
((c1: primitive state Inside-zone(p1, z1))
(c2: primitive state Inside-zone(p2, z1)))
constraints:
(c2 during c1)

Motion tracking and object identification was done as described previously. Mid-level primitive actions were generated using the tracking information. Sample images from the the dataset showing passengers disembarking are shown in figure 3. We manually extracted a few segments containing some activity, and automatically annotated the video with five different tags. The results of the annotation are shown in table 10. Most of the missed detections were due to errors in low-level detection and tracking modules.

Activity	Total	Correct	Missed	False
Passenger Embarkation	25	22	3	1
Passenger Disembarkation	25	21	4	0
Aircraft Arrival	2	2	0	0
Aircraft Departure	1	1	0	0
Luggage Cart Activity	5	4	1	0

Table 10: Detection results on the TSA airport surveillance dataset.

Figure 3: Sample Images from TSA dataset showing passengers disembarking.

5. CONCLUSIONS

We have discussed ontology design for human activities in surveillance settings. Building upon current ad-hoc ontology designs, we have presented a more systematic approach and have have shown quantitatively the improvement in recognition performance on two real-world surveillance domains.

Acknowledgments: This research was funded in part by the US government VACE program.

6. REFERENCES

[1] D. Chen, J. Yang, and H. Wactlar. Towards Automatic Analysis of Social Interaction Patterns in a Nursing Home Environment from Video. In *MIR Š04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 283–290. ACM Press, 2004.

[2] B. Georis, M. Maziere, F. Bremond, and M. Thonnat. A Video Interpretation Platform Applied to Bank Agency Monitoring. In *IDSS'04 - 2nd Workshop on Intelligent Distributed Surveillance Systems*, FEB 23 2004.

[3] T. R. Gruber. Toward Principles for the Design of Ontologies used for Knowledge Sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928, 1995.

[4] A. Hakeem and M. Shah. Ontology and Taxonomy Collaborated Framework for Meeting Classification. In *ICPR (4)*, pages 219–222, 2004.

[5] J. Hobbs, R. Nevatia, and B. Bolles. An Ontology for Video Event Representation. In *IEEE Workshop on Event Detection and Recognition*, 2004.

[6] J. R. Hobbs and F. Pan. An ontology of time for the semantic web. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(1):66–85, 2004.

[7] S. Hongeng and R. Nevatia. Multi-Agent Event Recognition. In *ICCV*, pages 84–93, 2001.

[8] T. Huang, D. Koller, J. Malik, G. Ogasawara, B. Rao, S. Russell, and J. Weber. Automatic Symbolic Traffic Scene Analysis Using Belief Networks. In *AAAI*, pages 966–972, 1994.

[9] Y. Ivanov and A. Bobick. Recognition of Visual Activities and Interactions by Stochastic Parsing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):852–872, 2000.

[10] F. Vu, V. T.and Bremond and M. Thonnat. Temporal Constraints for Video Interpretation. In *15th European Conference on Artificial Intelligence*, Lyon, FRANCE, 2002.