Action: A Framework for Semantic Annotation of Events in Video

Melanie Feinberg and Ryan Shaw

School of Information Management and Systems, U.C. Berkeley
feinberg@alumni.sims.berkeley.edu; ryanshaw@sims.berkeley.edu

Abstract. We propose a model for semantic annotation of *events,* such as weddings or birthday parties, as depicted in video. Our framework consists of an event taxonomy, implemented as a faceted classification, and an event partonomy, implemented using the ABC ontology proposed by Lagoze and Hunter [1]. Our approach enables the annotation of a low-level physical action depicted in video, such as a kiss, to be linked to its higher-level event context (such as the kiss that signifies the conclusion of a Western wedding ceremony).

1. Introduction

This paper describes an attempt to develop a semantically rich model for annotating events in video. Taking our cue from cognitive psychology research on event perception, we use a combination of taxonomy and partonomy for our event annotation model. We also take advantage of the faceted classification structure from information science to enable robust querying and differentiation of similar events without specifying all event possibilities in advance. Our original taxonomy enables discrimination of events on seven key levels (facets). The facet structure both facilitates fine-grained distinctions between events and enables recognition of broad commonalities. Finally, we use a multi-layered partonomy, familiar from artificial intelligence, that uses the existing ABC ontology [1] for expression as RDF. Our partonomic structure relies on principles from cognitive psychology research to segment events into logical, recognizable parts.

2 Related Work

We cite work in multiple disciplines, which reflects our synthetic approach to this project. By assimilating principles from cognitive psychology, information science, and artificial intelligence, we can create a cohesive model for event annotation.

Our approach is grounded in the work of the cognitive psychologists Jeff Zacks and Barbara Tversky [2], who assert that people perceive events similarly to the way that they perceive objects. Zacks and Tversky assert that, like objects, events are perceived according to two sorts of hierarchical structures. Events are structured taxonomically (that is, with superordinate, basic, and subordinate categories, as initially described by Eleanor Rosch [3]) and partonomically (divided into salient parts, as described by Tversky and Barbara Hemenway [4]). Zacks, Tversky, and Iyer [5] conducted experiments to show that test subjects viewing videotaped events segmented the events in predictable, regular ways.

In our framework, the taxonomic part of the annotation clarifies an event in relation to other types of events (for example, weddings and birthday parties are both celebrations, while basketball is a sport). In implementing our event taxonomy, we used a *faceted* structure, a form that comes from bibliographic classification [6]. The ability to create new terms through combination is a particular advantage of faceted classification. All concepts do not need to be predefined, as new concepts can be created by combining terms from different facets. In addition, the ontology itself can be simpler and less redundant.

While faceted classifications have not yet been commonly used to describe events, the AI community has used partonomies to do so. Marvin Minsky's frames [7], Schank and Abelson's scenes and scripts [8], and Ortony and Rumelhart's event schemata [9] are examples of events being described in terms of their typical parts.

3. Ontology Structure

In this section, we describe our taxonomy and our partonomy.

3.1 Taxonomic Structure

We designed our taxonomy to include the following facets. Each facet identifies a separate set of descriptors, organized in a hierarchy from general to more specific. In classifying an event, descriptors can be chosen from some or all of the facets.

- Time (with sub-facets Boundaries, Ordering, Recurrence, and Duration). The Time facet includes descriptors to specify temporal aspects of an event, such as whether the event has strict beginning and ending points, whether event segments can be reordered, whether the event is part of a series, and variability in the event's total extent.
- Physical Effect (with sub-facets Product and State Change). This facet describes changes in the environment as a result of the event, whether that change involves the creation of a new product (such as baking cookies) or changes to an existing object (such as repairing a clock).
- Focus. This facet differentiates between events with identifiable focal points and those without. A focal point describes an element that, if not viewed, would compromise the sense of having seen the event. For example, video of a birthday party without showing the candles being blown out would seem incomplete.
- Organization. This facet describes the differences between events that have imposed structure and those that are more improvisational. For example, this facet seeks to describe the difference between a professional basketball game and a pickup basketball game on a public neighborhood court.
- Style. This facet indicates manner. For example, a birthday celebration in the United States is structured differently from one in Mexico.
- Activity. This is the basic descriptor. Expressing the activity generically allows for subtleties to be conveyed using the Purpose facet. For example, for the activity of playing music, context could further define the event as a performance, practice, audition, and so on. These latter distinctions, which might apply to many activities, are moved into the Purpose facet, reducing redundancy in the taxonomy.
- Purpose. This facet adds a more complex semantic layer onto the generic description enabled by the Activity facet. The Purpose facet differentiates playing the piano (with no additional purpose) from a piano competition or piano concert, for example.

The faceted structure enables us to differentiate between events that are similar in some ways but different in others, without explicitly specifying each possible variation. The faceted structure also enables us to describe events that might be unique or are impossible to anticipate, such as a birdcage-making contest. Such an activity might occur only once in the world, but we can specify this improbable event easily by combining descriptors from different facets. We might use an Activity facet descriptor to represent carpentry, a Physical Effect descriptor to indicate the product of a birdcage, and the Purpose facet to clarify a competition.

The use of facets makes searching for related footage more robust, as the search relates to concepts, and not keywords. For the birdcage example, one could search for video of carpentry, the Activity facet descriptor, without searching on additional facets, and obtain footage of any object being created through carpentry, not just birdcages. Similarly, one could search for competition, the Purpose facet descriptor, and obtain results of math competitions, swimming races, and eating contests in addition to the birdcage-making competition. And of course, one could use all three facets for specific results.

3.2 Partonomic Structure

The partonomic aspect of our framework describes the structure of individual events from the taxonomy, including sub-events, actions, agents, and objects, and how they relate to one another. In creating the partonomy we used strategies for event segmentation hypothesized by Zacks, Tversky, and Iyer [5].

The primary activity level represents the basic modules of the event. The generic action level represents the basic actions within each activity. At the specific action level, we indicate the different actions required for classes of variables that are involved in implementing a generic action. For example, obtaining refreshments, a generic action, differs if the guest is obtaining a beverage or a solid food. At the atomic action level, we specify the physical actions necessary to complete a generic action for the instantiation of a specific variable.

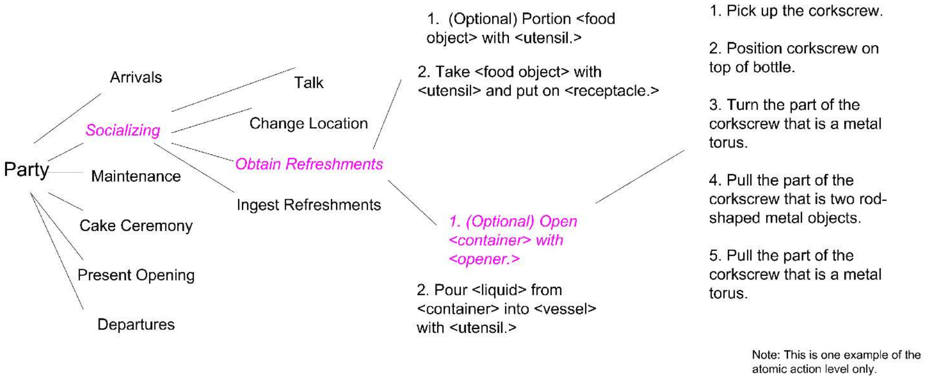
Party
Arrivals
Socializing
Maintenance
Cake Ceremony
Present Opening
Departures
Talk
Change Location
Obtain Refreshments
Ingest Refreshments
1. (Optional) Portion <food object> with <utensil.>
2. Take <food object> with <utensil> and put on <receptacle.>
1. (Optional) Open <container> with <opener.>
2. Pour <liquid> from <container> into <vessel> with <utensil.>
1. Pick up the corkscrew.
2. Position corkscrew on top of bottle.
3. Turn the part of the corkscrew that is a metal torus.
4. Pull the part of the corkscrew that is two rod-shaped metal objects.
5. Pull the part of the corkscrew that is a metal torus.
Note: This is one example of the atomic action level only.

Fig. 1. From left to right: primary activities, generic actions, specific actions, atomic actions.

Explicitly linking actions from different levels of description potentially allows for greater recall when searching for annotated video content. For example, a query for “birthday party” could be expanded to include sub-events such as “gift-opening.”Likewise, queries made at a more specific level of description can be expanded to return footage that has been annotated at a broader level.

4. Implementation

After conceptualizing our taxonomy and partonomy in proof-of-concept form, we formally expressed each of them as RDF graphs [10,11] and linked them together, as shown in Figure 2. For our taxonomy we took advantage of the Simple Knowledge Organization System (SKOS) Core, an RDF vocabulary developed for thesauri [12], while for the partonomy we utilized the aforementioned ABC Ontology [1]. The results can be browsed interactively at [13].

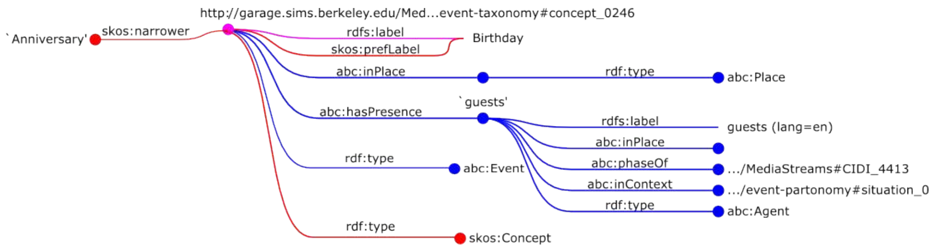

Fig. 2. An excerpt of the graph showing how the taxonomy (red) links to the partonomy (blue).

5. Conclusions

Video content is difficult to search. Video annotation can help by identifying and contextualizing video content at a level relevant to users' experience. Video of events is particularly in need of contextualized annotation, because the physical actions depicted in a particular video segment may reappear in many different contexts. To enable robust search and retrieval of video events, we need a multi-layered annotation framework that combines the low-level actions that facilitate maximum reuse with the higher levels that people are more likely to identify. To accomplish this goal, we have combined an event taxonomy, which classifies events in relation to similar events, with an event partonomy, in which events are successively segmented into smaller and smaller parts. In the future, we hope to use this conceptual model as the basis for a Semantic Web application that enables collaborative annotation of events depicted in web video.

References

1. Lagoze, C., Hunter, J. (2001) "The ABC Ontology and Model." *Journal of Digital Information* 2(2).
2. Zacks, J. M., Tversky, B. (2001) “Event Structure in Perception and Conception.” *Psychological Bulletin,* 127, 3-21.
3. Rosch, E. Principles of Categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and Categorization* (pp. 27-48). (Hillsdale, NJ: Lawrence ErlbaumAssociates, 1978).
4. Tversky, B., Hemenway, K. (1984) Objects, Parts, and Categories. *Journal of Experimental Psychology: General,* 113, 169-193.
5. Zacks, J. M., Tversky, B., Iyer, G. (2001) “Perceiving, Remembering, and Communicating Structure in Events.” *Journal of Experimental Psychology: General,* 130, 29-58.
6. Broughton, V.,“Faceted Classification as a Basis for Knowledge Organization in a Digital Environment: The Bliss Bibliographic Classification as a Model for Vocabulary Management and the Creation of Multidimensional Knowledge Structures,” *The New Review of Hypermedia and Multimedia* 7, no. 1 (2000): 67-102.
7. Minsky, M. “A Framework for Representing Knowledge.” MIT-AI Laboratory Memo 306, June, 1974.
8. Schank, R., Abelson, R. *Scripts, Plans, Goals, and Understanding. An Inquiry into Human Knowledge Structures.* (Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1977.)
9. Rumelhart, D. E., Ortony, A. (1977) “The Representation of Knowledge in Memory.” In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the Acquisition of Knowledge* (pp. 97-135). Hillsdale, New Jersey: Lawrence Erlbaum Associates.
10. Action Taxonomy, http://www.sims.berkeley.edu/~ryanshaw/action/taxonomy.rdf.
11. Action Partonomy, http://www.sims.berkeley.edu/~ryanshaw/action/partonomy.rdf.
12. Alistair J. Miles, Nikki Rogers, and Dave Beckett, “SKOS-Core 1.0 Guide,” SWAD-Europe, http://www.w3.org/2001/sw/Europe/reports/thes/1.0/guide/.
13. Action RDF Visualization, http://dream.sims.berkeley.edu:8080/ryanshaw/visualize.