Assessing Behavioral and Computational Approaches to Naturalistic Action Segmentation

Meredith Meyer[1] (mermeyer@umich.edu), Philip DeCamp[2] (decamp@media.mit.edu), Bridgette Hard[3] (martin@psych.stanford.edu), Dare Baldwin[4] (baldwin@uoregon.edu), Deb Roy[2] (dkroy@media.mit.edu)

[1]Department of Psychology, University of Michigan, Ann Arbor, MI 48103 USA
[2]Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139 USA
[3]Department of Psychology, Stanford University, Stanford, CA 94305 USA
[4]Department of Psychology, University of Oregon, Eugene, OR 97403 USA

Abstract

Recognizing where one action ends and another begins is an automatic and seemingly effortless process that supports understanding of goal-directed action. One characteristic of such action segmentation is that it is hierarchical; it reflects the goals and sub-goals of an actor, which correspond to coarse- and fine-grained action units respectively. We report on the success of one method of assessing hierarchical segmentation of naturalistic footage taken from an extensive corpus of unscripted human action (Speechome project, e.g., Roy et al., 2006). Results indicate that hierarchical segmentation occurs in an on-line fashion, with event boundaries marked by surges in attention that are modulated based on whether a boundary marks a fine, intermediate, or coarse unit. We also describe a method by which objective changes in an actor's movement can be measured and analyzed as a predictor of participants' segmentation behaviors.

Keywords: action segmentation; event processing

Drawing inferences and generating predictions about others' actions are processes most people undertake every day. The ways in which people use such inferences and predictions to make sense of others' action is supported in part by the ability to segment continuous action into discrete units. For instance, while observing an individual preparing dinner, we might identify and recognize individual units of action such as chopping a carrot, opening a refrigerator, or rinsing off a dish. Investigations of action segmentation have suggested that people are highly consistent in where they judge event boundaries to exist; people typically report dynamic human action to consist of units corresponding to initiation or completion of goals, with considerable agreement across individuals regarding where event boundaries are located (Baldwin & Baird, 1999; Newtson, Engquist, & Bois, 1977; Zacks, Tversky, & Iyer, 2001). Further, action segmentation is seemingly spontaneous and automatic, engaged in as a routine and ongoing component of perception (Hard, 2006; Zacks & Swallow, 2007).

The apparent ease with which people recognize breakpoints in action is remarkable given the complexity of the action stream itself. Human action is unquestionably a rich and highly variable stimulus; it is evanescent, often proceeds without pauses to mark the completion of individual units, and frequently features occlusion of relevant objects and body parts. Further, the underlying structure of action is also complex, typically characterized by a hierarchy reflecting the goals and sub-goals of an actor (e.g., Schank & Abelson, 1977).

Notably, human observers' skill in segmenting the action stream has been observed on a variety of different levels in line with this hierarchical structure. For example, segmentation of "chop carrot" can be on a *coarse* level, with event boundaries noted at the onset and offset of the entire chopping event, or it can be on a *fine* level, with each vertical movement of the knife noted as marking a discrete unit. In tasks assessing hierarchical segmentation, here again a high degree of consistency has been observed in people's segmentation behaviors (e.g., Hard, 2006; Zacks et al., 2001a), and fMRI studies have revealed differing activation levels in frontal and posterior areas in response to fine and coarse event boundaries, suggesting that the distinction between fine and coarse units is psychologically real on a neural level (e.g., Zacks et al., 2001b).

The ability to determine when one action has ended and another has begun, as well as segmenting action on multiple levels, supports how we make sense of the goal-directed action we observe in others. The fact that hierarchical event segmentation appears to be a relatively effortless process despite the complexity of the action stream itself suggests the workings of an equally complex system enabling this segmentation. Of particular relevance for the current studies, work by Hard and colleague (e.g., Hard, 2006; Hard & Recchia, 2006) suggests that event boundaries are processed differently than within-unit moments, with the detection of boundaries associated with a transient increase in cognitive processing load.

The idea that event boundaries might elicit an upsurge in cognitive processing is consistent with a comprehensive account of action segmentation put forth by Zacks and colleagues. These authors (e.g., Kurby & Zacks, 2007; Zacks et al., 2007) describe the Event Segmentation Theory, an account of how the human observer perceives and conceptualizes action in terms of events. A crucial component of Event Segmentation Theory rests on the observer's ability to make predictions about upcoming action. Such prediction generation is considered a spontaneous, online process that integrates incoming sensory information with prior knowledge and learning in an attempt to create a stable "event model." Event units correspond to periods in which prediction error rate is low; the observed action is consistent with the predictions being

made by the perceptual system, and the event model is stable. For example, within the event of cleaning off plates at the kitchen sink, the predictive system is able to generate accurate predictions of further plate cleaning based on such cues as the person's movements and prior knowledge about kitchen clean-up. Event boundaries, in contrast, are experienced when prediction error rate is high; to extend the example above, such boundary moments are likely to occur at the completion of a task (e.g., cleaning off plates in the kitchen) and before the initiation of another task (e.g., wiping the countertop), because these moments correspond with a reduced ability to predict the onset and content of the second event.

In order to update the event model at moments of reduced predictability, the system is believed to increase attention to the perceptual characteristics of the action stream and to activate new event schemata to replace the prior unsuccessful one. Hard and colleague (Hard, 2006; Hard & Recchia, 2006) provided an empirical test of whether boundaries were indeed associated with differential degrees of cognitive processing. As their methodology formed the basis of the first experiment in the current study, an in-depth explanation of their methods is in order. These authors reasoned that well-known paradigms developed for investigations of hierarchical processing of text would also be suitable for revealing aspects of hierarchical processing of action. In one such text processing study, individuals saw one word at a time from a passage of text and advanced themselves through word-by-word by pressing a button. The length of time between button presses was the primary dependent variable in this "moving window" method, with the idea being that longer reading times would be indicative of increased cognitive load associated with integration of past elements within and across text units into comprehensible larger units. Results indicated that participants tended to spend longer periods of time on words located at the ends of unit boundaries. Further, this "wrap up" effect was modulated by the level of any given unit; reading times were longer for words located at the ends of clauses and longer still for words located at the ends of sentences (Haberlandt & Graesser, 1989).

To study processing of hierarchical action using a similar technique, Hard and colleague adapted the moving window method for use with human action by asking participants to advance through a sequence of still-frame images. These images were taken from regular time intervals of footage of scripted human goal-directed action (e.g., one still-frame image sampled every second). Following this "slideshow" viewing phase, participants watched the live action footage from which the still images had been sampled and marked with a button press the locations of action boundaries (hereafter, 'breakpoints'). Participants completed this segmentation task a total of three times, providing judgments on fine, intermediate, and coarse levels.

Results from the slideshow task indicated that participants tended to spend a longer period of time looking at images close in time to moments judged to be breakpoints in comparison to images taken from within action units, suggesting that breakpoints elicited surges in attention. Further, paralleling results observed in text processing, the effect was modulated by the level of the action breakpoint, with slides close in time to moments judged as coarse-grained breakpoints receiving the longest looking times and those near fine-grained breakpoints receiving the least. This phenomenon, dubbed the dwell time effect, provided evidence that hierarchical segmentation occurs as part of real-time perception, without requiring explicit after-the-fact judgments of breakpoint locations. It further demonstrated the cognitive importance of action breakpoints; heightened attention was associated with moments participants explicitly judged to be breakpoints, and this effect was modulated based on whether that breakpoint was judged to be coarse, intermediate, or fine.

In the current paper, we report on another study that investigated hierarchical processing of action, this time using in vivo recordings collected from the Human Speechome Project. Audio-video data was collected from from the home of a single child using 11 ceiling mounted cameras and 16 boundary layer microphones. Over the first three years of the child's life, 90,000 hours of video was collected, representing roughly 70% of the child's waking experience (Roy et al., 2006).

As described above, past work has made much progress on elucidating the cognitive processes that make up the system enabling segmentation; however, these studies have examined segmentation of either scripted or animated scenes (e.g., Hard, 2006; Hard & Recchia, 2006; Zacks, 2004; Zacks et al., 2001a; Zacks, Kumar, & Abrams, 2009). The use of Speechome footage has the advantage of providing unscripted activity, allowing a test of the validity of methods that have been successful in revealing aspects of hierarchical segmentation of more artificial action scenes. Validation of the dwell time paradigm in Speechome footage additionally provides opportunities for the assessment of automated means of detecting action units, the topic taken up in Study 2.

Study 1 Method

Stimuli

Images for a slideshow viewing task were created by extracting one image every second from a 108-second movie clip take from the Speechome corpus (e.g., see Figure 1). The clip selected depicts an adult male preparing a meal. This video clip also served as the live action footage for which participants provided explicit segmentation judgments. For the explicit segmentation task, a different, 40-second clip of a woman cleaning the kitchen was used for training purposes.

Participants and Procedure

Participants were 28 university students (14 male) receiving class credit for participation. The experiment had two major phases, the *slideshow viewing task* and the

Figure 1: Sample image from slideshow depicting a person preparing food.

explicit segmentation judgment task. All participants began the session with the slideshow viewing task, in which they were instructed to advance at their own pace through the 108 still-frame images. Participants were told to click a mouse to advance the pictures. A Macintosh G4 computer was used to present stimuli on a 19.5" x 12" monitor, and Psychtoolbox (Brainard, 1997) was used to record participants' responses.

Following the slideshow, participants heard a brief description of how action can be seen as consisting of units, and examples of fine, intermediate, and coarse units in actions unrelated to those displayed during test were provided in these instructions. Participants then provided explicit judgments of where they believed breakpoints to be located, first providing judgments for the training video and then for the 108-second test (Speechome) video. Participants indicated their judgments with a key press. Participants were asked to provide segmentation judgments on fine, intermediate, and coarse levels, resulting in a total of three viewings of the movie clip. Half of the participants were asked to segment on a fine level on their first viewing of the clips, followed by segmenting on an intermediate level, and finishing with segmenting on a coarse level (fine-to-coarse order). The other half was asked to segment in the reverse order (coarse-to-fine order). Assignment of participants to these orders was random.

Study 1 Results

Do participants' explicit segmentation judgments reflect understanding of hierarchical structure?

One important preliminary question to answer is whether participants understood our instructions regarding segmentation on fine, intermediate, and coarse levels. Because we planned to compare the dwell times provided by each subject to their explicit breakpoint judgments made afterwards, it was important to ensure that participants differentiated among fine-, intermediate-, and coarse-level breakpoints during the explicit segmentation task.

Evidence for this understanding comes in part from results indicating that participants provided significantly different numbers of judgments for breakpoints at different levels, with fine-level breakpoints receiving the most judgments (M fine = 39.04 [SD = 23.32]), intermediate-level breakpoints receiving the next most (M intermediate = 12.68 [SD = 8 86]), and coarse-level breakpoints receiving the least (M coarse = 5.75 [SD = 2.81]), F (1.13, 30.42) = 61.44, $p < .0001$. (Greenhouse-Geisser statistics are reported due to violations in sphericity.) A significant linear trend characterized these data, F (1, 27) = 64.18, $p < .0001$. Thus, participants were clearly capable of recognizing breakpoints on different levels, providing the predicted differences in number of judgments according to level (fine vs. intermediate vs. coarse). As well, although individual differences in number of judgments were substantial (particularly in fine and intermediate judgments, as evidenced by the large standard deviations), 100% of participants provided the most judgments for fine breakpoints and the least for coarse breakpoints (sig. by a binomial test, $p < .0001$).

Participants were also fairly consistent in where they marked the locations of breakpoints. Figure 2 displays the number of fine, intermediate, and coarse level judgments across the 108 seconds of footage, with judgments "binned" into one-second intervals. As demonstrated by the distinct peaks and valleys reflecting moments commonly judged and rarely judged as breakpoints, respectively, it is apparent that participants frequently marked the same moments for all three levels of judgments, a pattern largely consistent with past studies using the same explicit segmentation method (e.g., Hard, 2006; Zacks et al., 2001a; Zacks et al., 2009).

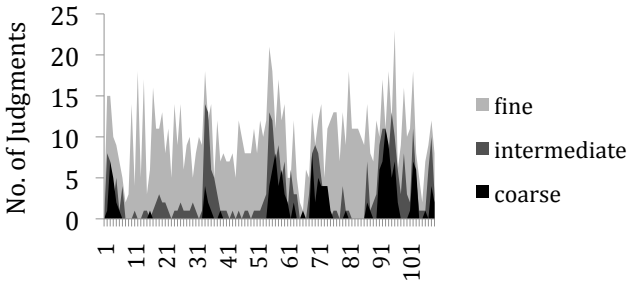

Figure 2: Participants' explicit judgments of fine, intermediate, and coarse level boundaries.

Does dwell time increase at breakpoints?

We next turned to one of the major hypotheses guiding Study 1, namely that participants' dwell time would be longer for images judged to be breakpoints compared to those that weren't. We used the participants' own explicit segmentation judgments, provided during the segmentation task, as the basis for determining which slides were considered breakpoints. Specifically, we applied a binning method, splitting the 108-second test clip into 1 second intervals, each corresponding to a single slide. Breakpoint judgments that fell into a given interval were matched to the corresponding slide, allowing us to classify breakpoint vs. non-breakpoint slides for each participant.

We then treated participants' raw dwell times to individual slides according to the following steps. Outliers (>3 SD above an individual's mean dwell time to all 108 slides) were removed from the data. Data were positively skewed, and thus a log transformation was applied. Due to

participants' tendency to dwell longer on slides at the beginning of the sequence and to speed up as the task continued, most participants' data were consistent with a power function. Significant portions of the variance were accounted for by the model for all participants (highest p value was .02). Thus, data were de-trended, and the residuals calculated based on the power function were used for analysis.

Because there were unequal numbers of slides in the different classifications (e.g., far fewer slides classified as breakpoints vs. non-breakpoints), means for each type were divided by standard deviations of that type, producing an effect size. All reported analyses are on these scores, hereafter referred to as dwell time scores. (Note that dwell time scores can be zero or negative since the residuals represent the difference between actual dwell time and times predicted by the power function; however, it is still the case that higher dwell time scores indicate overall longer dwelling on any given slide.)

A 2 (breakpoint status: breakpoint vs. non breakpoint) x 2 (segmentation order: fine-to-coarse vs. coarse-to-fine) mixed ANOVA (with breakpoint status as a within-subjects variable and segmentation order as a between-subjects variable) revealed only the predicted breakpoint status effect. Dwell time scores for breakpoint slides (M = .124, SEM = .046) were higher than for non-breakpoint (within-unit) slides (M = -.044, SEM = .026), F (1,26) = 6.40, p = .02. The main effect for segmentation order was not significant (M fine-to-coarse = .01, SEM .03; M coarse-to-fine = .07, SEM = .02), F (1, 26) = 3.1, p > .05, nor was the segmentation order x breakpoint status interaction significant, F (1, 26) = .03, p > .05. Dwell time scores were thus higher for breakpoints than non-breakpoints, supporting the first hypothesis.

Do dwell times vary according to fine, intermediate, and coarse levels?

Using the same binning method used to distinguish between breakpoint and non-breakpoint slides for each participant, classification of slides as breakpoints vs. non-breakpoints for each individual participant, slides were additionally categorized as falling at fine, intermediate, and coarse level boundaries. We then examined whether the dwell time effect was modulated based on whether a breakpoint was judged to be on a fine, intermediate, or coarse level. A 3 (segmentation level: fine, coarse, intermediate) x 2 (order: fine-to-coarse vs. coarse-to-fine) mixed between-within ANOVA was run, with segmentation level as the within-subjects variable and order as the between-subjects variable. Because of sphericity violations, we report Greenhouse-Geisser statistics. The predicted main effect for segmentation level was found, F (1.52, 39.43) = 16.17, p < .0001 (see Figure 3 for means). These differences were characterized by a significant linear trend, F (1, 26) = 21.20, p < .0001, with coarse-level breakpoints receiving the longest dwell times, intermediate-level breakpoints receiving the next longest, and fine-level breakpoints receiving the shortest dwell-times. The main effect for order was not significant (M coarse-to-fine = .161, SEM = .057; M fine-to-coarse = .087, SEM = .073), F (1, 26) = 1.23, p > .05; there also was no order x segmentation level significant interaction (F (1.57, 39.43) = .95, p > .05.

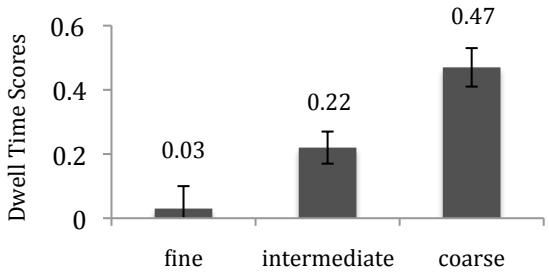

Figure 3: Dwell-time scores to slides designated as fine, intermediate, and coarse breakpoint. Data were characterized by a linear trend, p < .0001.

Study 2

Another line of investigation in action segmentation has focused on determining what perceptible features in the movement stream are relevant to segmentation. For instance, in the same study in which Hard and Recchia (2006) showed attentional differences to event boundaries, they additionally found that greater body movements on the part of the actor (as measured by overall pixel change between slides) significantly predicted observers' segmentation behavior. Similarly, in Zacks and colleagues' (2009) investigation of live action, the authors studied how changes in movement features such as the actor's acceleration and speed were predictive of observers' explicit segmentation judgments. In that study, an actor wore magnetic tracking devices on his hands while filming an action sequence, allowing for later extraction and calculation of the relevant movement features. The authors found that several movement features, including speed, acceleration, and change in distances among the actor's hands and head were predictive of observers' segmentation judgments, particularly for fine-grained event markings (see also Zacks 2004 for similar analyses with animated figures).

The ability to predict event boundaries based on perceptible features that can be extracted from video has great relevance to designers of informational systems that use identified actions as units of analysis. In addition to testing the validity of the dwell-time methodologies in naturalistic action, another goal of the current paper was to assess whether features visible in the action input were predictive of individuals' segmentation judgments. In Study 2, we extracted a set of predictive features, then analyzed how well these predictors correlated to the human judgments collected for Study 1

Study 2 Method

A set of motion features was extracted from the Speechome test clip using an accurate, semi-automatic

tracking system to annotate the positions of the body and hands of the actor appearing in the video (DeCamp & Roy, 2009). Positions were recorded as image coordinates (2D positions on the image, as compared to 3D positions in real space). Body position was defined as the center of the visible portion of the actor's head and torso. The positions of the hands were defined relative to the position of the body in order to reduce the covariance between them. After the position information was collected from the test video, it was used to compute the speed and acceleration of each body part, resulting in six features (see Table 1). The first and last seconds of data were also removed from analysis at this point because it was not possible to robustly define speed and acceleration at these points.

Kernel density estimation was applied to the breakpoints at each granularity level (i.e., fine, intermediate, and coarse). While this process smoothed the data, it also provided a continuous distribution of the breakpoints over time, which was more convenient for analysis than the raw judgment counts. Density estimation was performed with a Gaussian kernel. Bandwidths at were selected for each level using unbiased cross-validation, resulting in 0.92 s for fine breakpoints, 1.13 s for intermediate, and 1.27 s for coarse.

Study 2 Results

We found that each of the six features was significantly correlated to each breakpoint distribution (all p's < .001, see Table 1). The body speed feature achieved the highest correlation (r = 0.71) when correlated with coarse-grained judgments (see Figure 4). Right and left hand speeds had maximum correlations of 0.64 and 0.35, respectively. The acceleration features performed slightly worse, but were nevertheless significant.

Table 1: Correlations Between Visual Features and Breakpoint Distribution

	Correlation		
	Fine	Intermed	Coarse
Body Speed	0.49	0.65	0.71
Right-Hand Speed	0.47	0.64	0.64
Left-Hand Speed	0.45	0.44	0.35
Body Accel	0.40	0.52	0.54
Right-Hand Accel	0.35	0.51	0.47
Left-Hand Accel	0.34	0.40	0.36

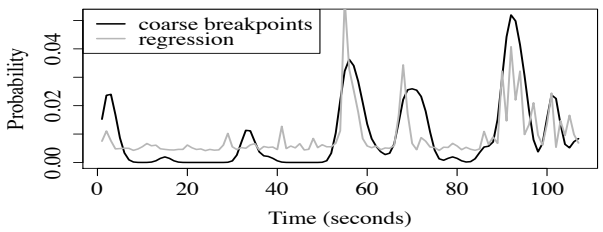

Figure 4: Univariate linear regression on coarse breakpoint distribution using body speed as predictor.

Discussion

In Study 1, we examined human observers' segmentation of naturalistic action, taking our stimuli from a large corpus of unscripted action (Speechome, e.g., Roy, 2006). Participants tended to dwell on images depicting breakpoints longer than non-breakpoints, and this difference was modulated based on whether a breakpoint was judged to be marking the completion of a fine-, intermediate-, or coarse-level unit. Despite the fact that our stimuli depicted naturalistic action, as well as the fact that participants had a decidedly different viewpoint of the action sequence itself than past studies of action (i.e., a ceiling-mounted camera provided the stimuli, and thus participants saw the actor from above), we replicated past findings of the dwell time effect (e.g., Hard, 2006; Hard & Recchia, 2006). Our findings suggest that the dwell time effect is a robust and valid phenomenon, capable of providing another window into the cognitive processes underlying segmentation.

The fact that participants' implicit behavior (dwell time) was associated with their explicit segmentation judgments also offers an exciting direction for future research within the developmental domain. There is clear indication already that infants as young as nine months can segment an action stream, a remarkable finding given infants' relatively impoverished understanding of goals and intentions (e.g., Baldwin et al., 2001; Saylor et al., 2007). Although this work represents an important demonstration of infants' action processing skill, the adaptation of dwell time methodology to this population has the potential to further expand our understanding of the developmental trajectory characterizing the segmentation process. The looking time methods used in these past developmental studies were not suitable for discerning hierarchical processing; further, the work examining hierarchical processing in adults has largely relied on participants' explicit understanding of what constitutes fine, intermediate, and coarse units (e.g., Zacks et al., 2001a, 2001b; Zacks et al., 2009), a task that is clearly beyond the capacity of infants and young children. We are actively pursuing adapting dwell time techniques for use both with preverbal infants as well as young preschool-aged children (e.g., Meyer, Hard, & Baldwin, 2009), a methodological advance that will allow us to study hierarchical processing across the lifespan.

In Study 2, we examined how perceptible movement features predicted human observers' judgments. Our results demonstrated that specific sources of information (i.e., head and hand speed and acceleration) were significantly associated with participants' segmentation judgments. Our results are consistent with similar movement change analyses performed by Zacks et al. (2009), suggesting that analysis of movement features may have broad utility in the design of automated systems of action analysis.

Notably, we additionally observed results that differed from those of Zacks et al., (2009); whereas we observed lower correlations as the judgment granularity was increased (i.e., correlations were highest when examining coarse-grained judgments and lowest when examining fine-grained

judgments), Zacks and colleagues actually observed the opposite. We speculate that this might be attributed to the differences between videos; in our footage the actor had no discernible facial features, and local movements of the hands and fingers were difficult to see; this may have reduced the ability of subjects to identify breakpoints as consistently at finer granularities. As well, the actor in our video moved his entire body through space (e.g., walking from a kitchen island to the sink), whereas the actor in Zacks et al.'s videos was seated. These gross bodily movements were frequently judged as coarse breakpoints and were clearly associated with several of our movement cues. Finally, the use of 2D video annotations in place of 3D motion sensor features may have provided less accurate measures that limited our ability to predict finer-grain events. In any event, the differences we observe offer inviting topics for future investigation relevant to the development of automated action analysis.

To summarize, we both validated the dwell time effect in naturalistic stimuli as well as found objective movement parameters predictive of individuals' segmentation behavior. The latter finding is of great relevance for researchers developing automated action analysis systems. Given that tracking whole people is now feasible for many types of video, current tracking technologies may enable the first steps towards systems that can automatically segment and identify actions from raw video, opening up new possibilities for human behavioral analysis.

Human action is an undeniably rich and complex stimulus. Yet, as we parse the events of our daily lives with little thought or apparent effort, the process may strike us as trivially easy. Nevertheless, the complexity of human action is apparent upon any attempt at formalization, and it poses a considerable challenge towards understanding human cognition. In this paper, we supply part of the solution by demonstrating how the human mind reacts and imparts structure to action sequences as they unfold. We also provide promising results from attempts to predict and model these reactions, suggesting future possibilities for the data driven analysis of events at a massive scale.

Acknowledgements

This research was supported by the U.S. Office of Naval Research, award no. N000140910187.

References

Baldwin, D., & Baird, J. A. (1999). Action analysis: A gateway to intentional inference. In P. Rochat (Ed.), *Early social cognition*, (pp. 215–240). Hillsdale, NJ: Lawrence Erlbaum Associates.

Baldwin, D., Baird, J., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development, 72*, 708–718.

DeCamp, P., & Roy, D. (2009). Human-machine collaborative approach to tracking human movement in multi-camera video. *Proceedings of the 2009 International Conference on Content-based Image and Video Retrieval (CIVR)*.

Haberlandt, K., & Graesser, A. C. (1989). Processing of new arguments at clause boundaries. *Memory & Cognition, 17*, 186-193.

Hard, B. (2006). Reading the language of action: Hierarchical encoding of observed behavior. Doctoral dissertation, Stanford University.

Hard, B., & Recchia, G. (2006). Reading the language of action. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, (pp. 1433-1439), Vancouver, CA.

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences, 12,* 72-79.

Meyer, M., Hard, B., & Baldwin, D. (2009, October). Children's processing of action boundaries. Poster presented at Cognitive Development Society, San Antonio, TX.

Newtson, D., Engquist, G., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology, 35*, 847–862.

Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., Mavridis, N., Tellex, S., Salata, A., Guinness, J., Levit, M., & Gorniak, P. (2006). The human speechome project. *Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication.* (pp. 192-168).

Saylor, M. M., Baldwin, D., Baird, J. A., & LaBounty, J. (2007). Infants' on-line segmentation of dynamic human action. *Journal of Cognition and Development, 8,* 113-128.

Schank, R. C., & Abelson, R. P. (1977). Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Hillsdale, NJ: Lawrence Erlbaum Associates.

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science, 28*, 979–1008.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., et al. (2001b). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience, 4*, 651–655.

Zacks, J. M., Kumar, S., & Abrams, R. A. (2009). Using movement and intentions to understand human activity. *Cognition, 201,* 201-216.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind/brain perspective. *Psychological Bulletin, 133*, 273-293.

Zacks, J. M. & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science, 16,* 80-84.

Zacks, J. M., Tversky, B., & Iyer, G. (2001a). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General, 130*, 29–58.