Segmenting action for computers and humans: Possible links to intentional understanding*

S. Killingsworth, M. M. Saylor, and D. T. Levin

Psychology and Human Development
Vanderbilt University
Nashville, TN 37209, USA

stephenkillingsworth@yahoo.com
m.saylor@vanderbilt.edu
daniel.t.levin@vanderbilt.edu

Abstract — The ability to identify discrete actions within a continuous stream of motion interests researchers across a range of fields, including cognitive science, engineering and robotics. The present chapter reports on a study investigating adults’ segmentation of human action for a computer and a human. Consistent with observers using different levels of analysis—bottom-up versus top-down—for the two audiences, participants inserted more segments for the computer than for the human.

Index Terms — action analysis, segmentation, theory of mind, intentionality.

I. INTRODUCTION

A. General Introduction

Human action is radically complex: it is evanescent, continuous, and typically contains multiple actors, objects, and events. Such complexity raises questions about observers’ ability to extract structure from continuous human behavior. One set of questions concerns the nature of the units extracted from observed motions. Previous research has revealed a remarkable degree of consistency in the temporal locations at which infant and adult observers divide units of action; both groups reliably isolate units coinciding with actors’ goals and intentions [1], [2], [3], [4].

A second set of questions concerns possible mechanisms supporting human action analysis. Previous accounts have posited that top-down and bottom-up processes contribute to action parsing [5], [6]. Bottom-up processes draw on basic perceptual features of actions (e.g., the trajectory and velocity of motion) while top-down processes recruit conceptual information. Some types of conceptual information that may be recruited are: the physical properties of objects, forces acting on objects, social conventions, action schemas, and knowledge of others’ likely mental states, specifically, likely goals and intentions (see Reference [7] for a detailed discussion). Recent research suggests that adults use both top-down and bottom-up processing when segmenting action [8], but questions remain about how conceptual information influences segmentation. In this paper, we will focus specifically on the impact of intentional information.

In what follows, we begin with a brief review of previous research, suggesting great consistency in parsing strategies, to set the stage for a discussion of possible mechanisms supporting action segmentation. Then, we discuss how knowledge of others’ intentional capacities might influence action segmentation, before describing the present study.

B. Action Parsing in Infants and Adults

Several recent studies have highlighted the remarkable consistency with which infant and adult observers segment continuous human behavior into units. In each study, observers extract units that are commensurate with an intentional analysis of human action.

Saylor and colleagues have conducted two examples of such previous work [1], [3]. In the first study, 10-11-month-old infants revealed sensitivity to structure in human action by looking longer at events including interruptions of intention-relevant units than at events with interruptions occurring at intention boundaries [1]. Saylor has recently extended this finding by highlighting the scope and robustness of infants’ action analysis skills with a demonstration of 9-11-month-old infants’ detection of correspondences between endpoints of intention-relevant units and arbitrary tones [3]. Unlike the first investigation, this more recent study included six different action sequences that were somewhat novel to infants (the first study included only kitchen cleaning events). In addition, events in this second investigation were presented simultaneously in pairs, thus increasing the challenge of the segmentation task. Even still, infants perceived intentional structure in the action sequences. Taken together, these studies suggest that 9-11-month-old infants possess a robust set of skills for segmenting

* This work was partially supported by an NSF HSD grant (#4261003031) to Daniel Levin and Megan Saylor.
human action into units coinciding with actors’ goals and intentions.

Several studies have revealed similar findings with adults. In studies with adults, participants have been asked to explicitly indicate the end of an action segment (by generating a set of “break points”) and/or to describe action segments. First, Newson and colleagues revealed that adults converge on a similar set of break points when they are asked to segment a continuous action sequence such as motorcycle repair [9], [10]. In a more recent study, Baird and colleagues have revealed that such break points coincide with the intentions and goals observers’ attribute to actors [2]. In their first study, adults were presented with continuous action sequences and asked to identify points at which actors’ goals and intentions were successfully enacted. Participants did so with a high degree of consistency. Following this procedure, a second group of participants were shown the videos with tones placed either at the endpoints of intentions or in the midst of action. This group of participants showed reliably better memory for the position of endpoint tones than midpoint tones, suggesting that the transitions between actions have some special status in the analysis of human behavior (consistent with the research on infants described above).

Other researchers have probed adults’ organization of human action sequences. At issue is how adults relate distinct acts or events – once extracted – to one another in constructing and remembering complex action sequences. In particular, Zacks and colleagues have asked whether adults’ processing of dynamic human action displays hierarchical organization – with larger action units (e.g., putting a sheet on a bed) subsuming smaller action units (e.g., putting the top end of the sheet on the bed, putting the bottom end on the bed, smoothing the sheet)[4]. Participants in Zacks et al.’s study were asked to segment action sequences across a variety of tasks, including segmenting as an action unfolded, and describing segments from memory after actions were observed. In each of these tasks, participants’ judgments about the location of action endpoints revealed alignment between endpoints of small action units and endpoints of large action units. In other words, participants appeared to be embedding small action units within large action units.

C. Levels of Analysis

The previous research invites the question of how such action parsing proceeds. At least two possible mechanisms have been proposed: bottom–up and top-down processing. Top-down action processing involves expectations based on world knowledge. One aspect of such knowledge concerns reflection on the types of intentions an individual is likely to have in a particular context. On viewing someone making dinner or cleaning a closet, for example, observers would draw on their extensive knowledge of such activities and use this knowledge to segment the action into meaningful units (placing unit boundaries at the endpoints of distinct intentions). Bottom-up processing involves an identification of salient low-level patterns in the movement stream. Such patterns may include the configuration of motions that occur as an intention is initiated (e.g., body, eyes, and limbs moving toward goal objects) or statistical regularities in motions occurring together when initiating intentional acts [5], [6], [11].

Leaving aside the issue of how infants might segment continuous motion (though, there is a very good possibility they will draw on the same set of resources as adults when they have adequate knowledge in a domain), several authors have suggested that adults draw on both mechanisms when processing action, and recent findings corroborate their claims [5], [6], [8], [11].

D. How Knowledge of Intentional Agency Might Affect the Analysis of Action

One remaining question is how information about intentional agency affects observers’ analysis of action sequences. This question has been approached by either manipulating the perceptual features of an object or object motion, or by varying what observers are told about the agent(s). Several studies investigating infants’ analysis of action suggest that such knowledge influences action analysis even for the youngest observers, e.g., [12], [13], [14], [15], [16]. For example, clever studies by Susan Johnson and colleagues suggest that infants are willing to treat a broad class of (potentially) animate agents as intentional if they possess features characteristic of living things (e.g., contingent movement) [12], [13]. In addition, in separate studies Woodward and Meltzoff have revealed that infants analyze actions of human and mechanical agents (e.g., robotic arms) differently. In particular, they appear to attribute goals and intentions to human agents, but not to mechanical arms [15], [16]. This general finding of differential treatment of human and mechanical agents appears to extend to adults as well.

Among studies with adult participants, Levin and colleagues have demonstrated that adults’ attributions of perceptual (e.g., seeing) and cognitive capacities (e.g., memory) to mechanical systems differs according to how ‘human like’ (and potentially capable of intentional representations) the agents are. For example, in one study [17], participants were offered two descriptions of a computer – in one case the computer was described in anthropomorphic terms (e.g., with a typical proper name, using mentalistic verbs, and animate pronouns) in the other case the same system was described in mechanical terms (e.g., referred to via an acronym, without mentalistic verbs, and using the pronoun “it”). In the anthropomorphic condition, participants were more likely to predict errors in visual
processing and memory that other studies have found are attributed to humans, e.g., [17], [18]. This finding strongly suggests that adults modify their analysis of behavior according to the representational capacities of the agent (see Saylor and Levin chapter in the same volume for related findings with preschoolers).

In addition, a second recent study suggests that observers alter their segmentation strategies when an action sequence is described as having a non-intentional origin (i.e., randomly generated by a computer versus intentionally generated by people) or when an action sequence contains motions that are truly non-intentional (generated randomly by equations) as opposed to being generated by humans. In particular, participants inserted more units (consistent with bottom-up processing) and chose units that lined up with a stable set of movement features for the randomly generated motion [8]. Questions remain, however, about how adults' analysis of action might be influenced by the audience of an action sequence.

In the present study, we have manipulated the level of intentional agency of the audience for which one identifies meaningful units. This manipulation allows us to investigate the possibility that segmentation based on knowledge of familiar events can be altered by what is considered “meaningful” for a particular audience. Answers to this question may have implications for the design of mechanical agents that learn from human action and demonstrations. Additionally, this study looks specifically at whether observers are more likely to use a top-down mechanism when the audience is human (and therefore capable of intentional representations) and a bottom-up mechanism when the audience is a computer (and therefore not capable of intentional representations). Finally, we asked participants to report what information they used to segment the videos differently for the audiences and examined whether these strategies predicted differences in processing (based on the numbers of segments generated for the two audiences). Unlike previous investigations that have asked participants to segment action sequences at specific levels consistent with a top-down or bottom-up strategy [2], [4], [8], in the current study, we left the segmentation task open to enable investigation of whether spontaneous differences in observers’ segmentation strategies emerged with different audiences.

II. METHOD

A. Participants

Twenty-eight students from Vanderbilt University (male, n=8; female, n=20; ages 18-23, mean age=20) participated in the study for class credit or candy compensation. Two participants were excluded because the procedure was administered incorrectly and two participants were excluded for misunderstanding the instructions leaving twenty-three participants whose data were included in analyses.

B. Materials

The stimuli used in this study were two videotaped sequences of action used in previous studies to investigate action parsing in infants and adults [1], [2].

In the “Toy Room” video (duration: 54.7 s), a woman straightens a toy room. In doing so, she places a plastic tub on a shelf, hangs a shirt on a coat hook, and places some Lego blocks in a tub. In the “Kitchen” video (duration: 49.5 s), a woman cleans a glass, hangs a towel on an oven handle, returns ice cream to a freezer, and places a bowl in a dishwasher. In both videos, the action proceeded continuously without pauses between the actors’ individual motions. Certain “error” actions within the videos were repeated (2 in each video), with the actress appearing to have to make adjustments to complete an action. All stimuli were presented in the MacOS X operating system (monitor vertical refresh: 89Hz; the monitor measured 31.5cm x 23.5cm) using FinalCut Pro. The rectangle containing videos measured 9.2cm x 7.4cm.

C. Design and Procedure

Participants were seated approximately 60cm from the display. Participants were first instructed that they would be dividing videos into segments as the movie played and would then be given an opportunity to adjust their segments using a standard movie controller – allowing them to navigate to any point during the movie by scrolling. Participants were shown how to scroll through a movie and how to insert and delete segments using a sample movie clip displaying the text “Demo”.

Participants were informed that they would be segmenting videos for one of two audiences: a computer system with action analysis capabilities or a human viewer. They were asked to divide videos into segments that would be most easily identified by the audience for whom they were parsing (e.g., computer). After dividing the videos into segments for the first audience, they segmented the videos for the other audience (e.g., human). To prevent participants from considering both audiences at the same time, they were not given the instructions for the second audience until after having segmented both videos for the first audience.

The segmentation procedure was as follows: participants first watched a video without introducing segments to familiarize them with the events occurring in a clip. Next, during a real-time phase, participants divided the video into segments as the movie played. During a subsequent manual adjust phase, participants were given the opportunity to adjust the segments they had generated during the real-time trial, by adding, deleting or changing the position of segment markers. At the conclusion of the manual adjust trial, half of the participants were asked to write down a description of the content of their segments so that the audience could
learn to segment the videos. No differences in segmentation were observed between participants who described and those who did not so this manipulation will not be discussed further. We are in the process of coding the description data, so we will not discuss it in this chapter. The segmentation sequence was the same for both test videos in the two audience conditions. The order of video and audience presentation was counterbalanced across participants. After parsing for both audiences, some participants (n = 21) were asked what information they had used to segment the videos differently for the computer versus human audience. Our first 7 participants were given a less detailed questionnaire that did not include this question.

III. RESULTS

A. Number of Segments

To examine differences in how participants parsed videos for the two audiences, we conducted an analysis of variance (ANOVA) on the number of segments generated during each video viewing with audience (computer vs. human) and trial type (“real-time” vs. “manual adjust”) as repeated measures. Segments were used instead of breakpoints because some participants inserted breakpoints at the beginning or at the end of the videos, but most did not. No significant effects of description, movie, or orderings were found in preliminary analyses, so these terms were not analyzed further.

The main effect of audience was significant \[F(1,23) = 10.50, \ p < .01\], with the computer receiving significantly more parses (\(m = 7.91, \ SE = .56\)) than the human (\(m = 6.07, \ SE = .36\)). The main effect of trial type was also significant \[F(1,23) = 10.32, \ p < .01\], with the manual adjust trials receiving more parses (\(m = 7.28, \ SE = .44\)) than the real-time viewings (\(m = 6.70, \ SE = .33\)) overall.

The interaction between trial type and audience was also significant \[F(1,23) = 7.29, \ p < .05\]. There was very little difference in the number of segments between real-time and manual-adjust trials for the human audience whereas participants added parses during the manual adjust trials for the computer (see Fig. #1).

To examine the differences in the effect of trial type across the two audiences, paired samples t-tests were conducted for each audience. For the computer audience, the number of segments was significantly greater \[t(23) = -3.18, \ p < .01\] in the manual adjust trials (\(m = 8.40, \ SE = .67\)) than in the real-time trials (\(m = 7.42, \ SE = .48\)). For the human audience, the number of segments did not significantly differ between trial types \[t(23) = -1.37, \ p = .19\].

B. Questionnaire Responses

To analyze the effect of participants’ strategies on the number of segments generated, participants were grouped based on their responses to the strategy question on the post-experiment questionnaire (see Table #1). The majority of participants responded that they differentiated segments for the two audiences based on intentions or understanding an action’s meaning (the “intentions” group) – proposing that computers could not understand the meaning or intention in an action. Participants in the “Error Actions” strategy group reported that they had segmented the entirety of repeated (or “error”) actions for humans, but had split these segments for computers. The single participant in the “Descriptions” group was among those asked to describe segments for the audiences. This participant did not believe that the computer could easily identify segments based on descriptions. The one participant in the “Other” group reported that the closing segment would be understood by humans – but not computers – as a theatrical element. No significant effects or interactions were found for the strategy group factor in an ANOVA. In addition, when the participants who did not report having used repeated (or “error”) actions to differentiate the two audiences were removed from the analyses, the main effects reported in the above analysis of segment counts were still significant.

Additionally, correlations between the other post-experiment questions and numbers of segments were tested. A significant negative correlation was found between the number of segments generated for a computer system and subjective rating of how well computers can learn from prior experience and observation (\(r = -.475, \ p < .05\)). The correlation with

\[
\begin{array}{|c|c|}
\hline
\text{Strategy Group} & \text{Participants in Group} \\
\hline
\text{Intentions} & 8 \\
\text{Error Actions} & 3 \\
\text{Communicative Gestures} & 2 \\
\text{No Differences} & 1 \\
\text{Descriptions} & 1 \\
\text{Other} & 1 \\
\hline
\end{array}
\]

Fig. 1 Mean number of segments generated for each trial type for the computer and human audiences.
computers’ ability to learn from experience was significant for both the number of segments in real-time trials ($r = -0.460, p < 0.05$) and in manual adjust trials ($r = -0.468, p < 0.05$). There was also a significant negative correlation between the number of segments added in real-time trials for a computer and overall rating of computer intelligence ($r = -0.438, p < 0.05$). Finally, there was a significant positive correlation between the number of mathematics courses that a participant reported taking and the number or segments generated in computer real-time trials ($r = 0.476, p < 0.05$). No other correlations were significant.

IV. Conclusion

These findings suggest that adults vary action segmentation based on the audience for whom they are segmenting. Specifically, adults introduced more segment boundaries for a non-intentional computerized agent than for an intentional human agent. Furthermore, analysis of participants’ responses to questions about their ideas about computers’ abilities and intelligence revealed that segmentation was correlated with such views. Unlike previous research into how intentions affect the segmentation of action [8], our study investigated how imagined audiences influenced segmentation. When asked to identify meaningful units without reference to a specific audience, participants are likely to identify what is meaningful to themselves. Our findings suggest that when an audience different from the self is specified, participants identify units that are considered meaningful for that audience.

Our findings are consistent with a model of action analysis that involves a top-down analysis of the types of actions that would be meaningful for one’s audience (in this case, knowledge of whether the audience is capable of intentional agency). Participants may select from a range of models to apply to the actions. These include intentional and non-intentional models, but the options are not necessarily limited to these. For example, participants may select a non-intentional teleological mode of analysis in which the actions are mechanical, but designed to satisfy some human need. This initial decision would affect the degree to which participants are more likely to group actions based on the large units, reflecting intentions and goals, or on small units, reflecting representations of frequent changes in movement directions, velocities, and object-to-object distances.

An interesting question for further research is the degree to which participants’ initial attribution can be flexibly modified during the course of segmentation. On a relatively strict two-step hypothesis, participants make their initial attribution, and can use this to set task parameters that can guide attention during the task with relatively little further thought. Such an approach would allow the segmentation process to be responsive to contextual differences while retaining the effortlessness of a bottom-up perceptual process.

However, the cost would be a relative lack of responsiveness to signs of intentionality (or the lack of intentionality) that arise during the interaction. The alternative would be an ongoing analysis. This might occur if the initial intentional attribution requires effort to maintain, either because an explicit segmentation task such as ours is difficult and requires continuous cognitive monitoring, or because participants must sometimes overcome a default mode of action analysis. The current data provide some support for the latter because the human-computer difference was increased for manual adjust condition for the computer audience. Therefore, they may have started with a relatively easy default to focus on large intention-relevant units, and only with further effort decomposed these into component motions.

Several future directions arise from this study. First, are these different parsing strategies mirrored in participants’ verbal descriptions or action performances for humans versus computers? In descriptions, one might predict more use of mental state terminology for humans and a greater number of motion-related terms in descriptions for computers. Analyses of this sort are currently underway.

Second, though the audience for whom a participant was parsing clearly affected the number of segments generated, the role of intentionality in segmenting action is still in question. An additional way to examine the role of intentional agency would be comparing what occurred during segments generated for a computer to what occurred in segments for a human (currently in progress in our lab). From this analysis, several relationships could be examined. First, the percentage of segments for each audience that included completed intentional acts could be determined. Second, the correspondence between low-level motion cues and the locations of breakpoints could be examined – to determine the role of bottom-up processing for each audience. Finally, the analysis could be used to determine the degree to which a reported strategy actually directed how a participant divided the video.

When participants are asked to parse actions for two different audiences, it is possible that they develop strategies based on certain salient action features (like accidental actions – repetitions in action that occurred when the actor made an “error”) to distinguish between the perspectives. In the case of error actions, the strategy may be driven by intentional analysis. For example, accidental actions and action corrections may be seen as part of a single intention for humans. For a computer system, on the other hand, these may be parsed into separate actions.

Another factor that could contribute to the observed differences in segmentation is the audiences’ presumed level familiarity with the actions. Participants might recruit bottom-up mechanisms to segment actions that would be unfamiliar for an audience (because the likely
goals would not be clear). An unexpected possibility is that our participants were considering familiarity in addition to or instead of intentionality when segmenting actions for computers. Studies underway are addressing this by using an infant as the human audience: thus eliminating the difference in audiences' familiarity with the performed actions. This study will also enable an investigation of the role of perceived intelligence of the audiences.

In this first study we used a highly non-anthropomorphic computerized audience. An additional question is whether observers’ will adapt their segmentation strategies as it becomes more probable that an audience has intentional agency (e.g. a robot described in anthropomorphic terms as in [17]). Our initial findings suggest that people do so. Studies underway are addressing this issue.

The current findings have a range of implications for Human-Robot interaction. Among the most direct of these, people's attributions about the intentionality of different representational systems can affect their beliefs about what will be understood by those systems. Given the current state of computerized action and object analysis, our participants are probably correct in predicting that computers will revert to smaller units than people when perceiving actions. Accordingly, people may be able to produce demonstrations for computers that reflect this belief. However, people readily anthropomorphize computers, and this tendency can affect people's predictions about these systems' visual capabilities [17]. Therefore, people may or may not fail to match their demonstrations correctly for systems that vary in apparent intentionality. Given the recent emphasis on observational learning in robotics, we believe that research exploring people's beliefs about how these systems analyze actions will be important. Depending on how they are managed, these beliefs can either facilitate dynamic adaptive interaction or cause misunderstanding and frustration as people produce behaviors that are incomprehensible to robots and construe these systems' responses as annoying rather than helpful.

In summary, this research has demonstrated a good deal of flexibility in the parsing strategies employed by human observers—when they viewed an audience as incapable of intentional reasoning they segmented an action into smaller units than when an audience was viewed as being capable of intentional reasoning.

REFERENCES


ACKNOWLEDGMENT

The authors would like to thank the Vanderbilt Intentional Vision in Humans Robots and Computers group for lively discussions of the research and to Jodie Baird for access to her stimuli.