Cognition 177 (2018) 249–262

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

# The role of perspective in event segmentation

Khena M. Swallow*, Jovan T. Kemp[1], Ayse Candan Simsek

*Department of Psychology, Cornell University, 211 Uris Hall, Ithaca, NY 14850, USA*

ARTICLE INFO

*Keywords:*
Event segmentation
Perspective
Invariance
Event cognition

ABSTRACT

People divide their ongoing experience into meaningful events. This process, event segmentation, is strongly associated with visual input: when visual features change, people are more likely to segment. However, the nature of this relationship is unclear. Segmentation could be bound to specific visual features, such as actor posture. Or, it could be based on changes in the activity that are correlated with visual features. This study distinguished between these two possibilities by examining whether segmentation varies across first- and third-person perspectives. In two experiments, observers identified meaningful events in videos of actors performing everyday activities, such as eating breakfast or doing laundry. Each activity was simultaneously recorded from a first-person perspective and a third-person perspective. These videos presented identical activities but differed in their visual features. If segmentation is tightly bound to visual features then observers should identify different events in first- and third-person videos. In addition, the relationship between segmentation and visual features should remain unchanged. Neither prediction was supported. Though participants sometimes identified more events in first-person videos, the events they identified were mostly indistinguishable from those identified for third-person videos. In addition, the relationship between the video's visual features and segmentation changed across perspectives, further demonstrating a partial dissociation between segmentation and visual input. Event segmentation appears to be robust to large variations in sensory information as long as the content remains the same. Segmentation mechanisms appear to flexibly use sensory information to identify the structure of the underlying activity.

## 1. Introduction

The mind represents experience as a series of *events* that are organized into part-whole structures (DuBrow & Davachi, 2013; Kurby & Zacks, 2008). The process by which experience is divided into events, *event segmentation*, plays an important role in everything from language acquisition (Friend & Pace, 2011), to the recognition of other's intentions (Baird & Baldwin, 2001; Buchsbaum, Griffiths, Plunkett, Gopnik, & Baldwin, 2015), to episodic memory (Ezzyat & Davachi, 2010; Swallow, Zacks, & Abrams, 2009), and consequently to the ability to imagine future events (Buckner & Carroll, 2007). Despite its importance to cognition, the types of information observers use to divide continuous experience into meaningful events are underspecified. Though a close relationship between segmentation and visual input has been established (e.g., Hard, Recchia, & Tversky, 2011; Zacks, Kumar, Abrams, & Mehta, 2009), changes in the visual features of an experience are often correlated with changes in content (Cutting, 2014; Cutting, Brunick, & Candan, 2012).

This study disentangles the contributions of visual information and content to segmentation. It examines whether the same activity (content) is segmented differently when it is viewed from the actor's (*first-person*) perspective rather than from an observer's (*third-person*) perspective. These perspectives differ in their visual features and support differential access to the actor's goals, emotional state, and affordances with the environment (Jackson, Meltzoff, & Decety, 2006; Lamm, Batson, & Decety, 2007; Libby & Eibach, 2011; Nigro & Neisser, 1983; Storms, 1973; Taylor & Fiske, 1975; Vogeley & Fink, 2003; Vogt, Taylor, & Hopkins, 2003). In addition to testing whether segmentation is viewpoint dependent, contrasting segmentation across perspectives provides a unique window into the integration of sensory input with knowledge of events, and the ease with which observers can take an actor's perspective.

### 1.1. Event segmentation separates and organizes experiences

Event segmentation is measured by asking observers to view another person's activity (typically recorded on video). As they watch the activity, observers identify *event boundaries* by pressing a button

* Corresponding author.
*E-mail addresses:* kms424@cornell.edu (K.M. Swallow), jovan_kemp@brown.edu (J.T. Kemp), ac885@cornell.edu (A. Candan Simsek).
[1] Present address: Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, USA.

https://doi.org/10.1016/j.cognition.2018.04.019
Received 4 October 2017; Received in revised form 19 April 2018; Accepted 20 April 2018

whenever they believe one natural and meaningful unit of activity has ended and another has begun (Newtson, 1973). Despite the task's deliberate ambiguity, observers tend to perform it reliably, agreeing with themselves and with others about the timing of event boundaries (Newtson, 1973; Speer, Swallow, & Zacks, 2003). In doing so, they pick out moments in time that are important for perception and cognition. In the absence of a segmentation task, event boundaries are associated with increased activity in a network of brain regions (e.g., Zacks, Speer, Swallow, & Maley, 2010; Zacks, Tversky, & Iyer, 2001), impact memory for scenes and objects that were just encountered (DuBrow & Davachi, 2014; Ezzyat & Davachi, 2010; Newtson & Engquist, 1976; Radvansky & Copeland, 2006; Swallow et al., 2009), and may be sufficient for understanding an activity (Schwan & Garsoffky, 2004).

Observers are also sensitive to the hierarchical, part-whole structure of actions, even without an explicit segmentation task (Hard et al., 2011; Zacks et al., 2001). When asked, observers can vary the grain at which they segment an activity (Newtson, 1973) and identify events that capture parts of activities lasting several seconds to minutes. Shorter, *fine events* correspond more closely to individual actions performed on objects, while longer, *coarse events* correspond more closely to whole interactions with an object and actor goals (Zacks et al., 2001). As a result, fine events are often contained within coarse events (Hard et al., 2011; Zacks et al., 2001). Boundaries at both grains affect event processing during passive viewing tasks (Zacks et al., 2001).

### *1.2. The relationship between event segmentation, observer knowledge, and stimulus features*

Prominent models of event segmentation suggest central roles for both sensory information and knowledge of how experiences typically unfold. One model, event segmentation theory (EST; Reynolds, Zacks, & Braver, 2007; Zacks, Speer, Swallow, Braver, & Reynolds, 2007), is based on the idea that perception is fundamentally forward looking, that it is predictive. It claims that segmentation occurs when predictions no longer accurately capture the current situation, and that larger prediction errors produce boundaries between coarser grained events. According to EST, predictions are derived from semantic knowledge of types of events, actions, objects, and contexts and perceptual and sensory features of the current event.

The importance of bottom-up perceptual and sensory features for segmentation is strongly supported by research that demonstrates that the greater the change in the visual and auditory features of an activity (e.g., motion, body posture, location, scene, audio volume), the greater the likelihood that a boundary will be perceived (Cutting et al., 2012; Hard, Tversky, & Lang, 2006; Huff, Meitz, & Papenmeier, 2014; Magliano, Miller, & Zwaan, 2001; Magliano, Radvansky, Forsythe, & Copeland, 2014; Newtson, Engquist, & Bois, 1977; Sridharan, Levitin, Chafe, Berger, & Menon, 2007; Zacks, 2004; Zacks, Speer, & Reynolds, 2009; Zacks et al., 2010). Similarly, machine vision models often map local visual features (e.g., points in space-time with large luminance changes in the horizontal, vertical and temporal dimensions) to representations of action types (as in bag of words models, Peng, Wang, Wang, & Qiao, 2016). There are limits to the relationship, between visual changes and segmentation, however: Changes in an actor's clothing and large visual changes at film cuts do not increase the likelihood of segmentation on their own (Baker & Levin, 2015; Magliano & Zacks, 2011). Thus, segmentation is influenced by changes in a subset of observable features.

Though many acknowledge the importance of an observer's goals and knowledge in segmentation, establishing whether these factors work independently of sensory input is challenging. This is partly because changes in content are correlated with changes in visual features (Cutting, 2014; Cutting et al., 2012). For example, when an actor begins to empty her cart at a grocery store, changes in motion (the actor's movements), the spatial relationship between the actor and the cart (she moves to the side), and the actor's posture (she bends to pick up food) signal a change in the actor's goals. In the face of this relationship, most investigations of the role of knowledge and goals in segmentation have examined how changing an observer's knowledge affects segmentation. For example, learning statistical regularities in event sequences can lead observers to group smaller units into larger units (Avrahami & Kareev, 1994; Baldwin, Andersson, Saffran, & Meyer, 2008; Buchsbaum et al., 2015; Endress & Wood, 2011; Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013). In addition, segmentation behavior may change when an observer's knowledge of an actor, the actor's goals, or his or her activity changes (Bailey, Kurby, Giovannetti, & Zacks, 2013; Graziano, Moore, & Collins, 1988; Wilder, 1978; Zacks, 2004). There is also some evidence that changing the observer's task (e.g., from reproducing an activity, to judging traits) or beliefs about the purpose of the action may influence when observers segment an event (Cohen & Ebbesen, 1979; Massad, Hubbard, & Newtson, 1979). However, it is unclear whether these differences are greater than one might expect from measurement noise alone (cf. Speer et al., 2003). Outside the domain of segmentation there is also substantial evidence that an observer relies on motor and visual-perceptual knowledge to recognize and comprehend another person's actions (Blakemore & Decety, 2001; Fogassi et al., 2005; Glenberg & Kaschak, 2002; Stanfield & Zwaan, 2001; Wilson & Knoblich, 2005; but see Vannuscorps & Caramazza, 2016).

Other research contrasting the effects of knowledge and sensory input on segmentation suggest that, although knowledge may influence the grain at which events are segmented, boundaries are still identified when sensory input changes. For example, Hard et al. (2006) asked some participants to view animations five times before they segmented them. These participants rated the activities as more intentional and segmented them at a lower rate than participants who segmented the videos the first time they viewed them. Yet, both groups segmented the activities at similar points in time. Boundaries were also similar when the movies were played forward and backward. In all cases, increased visual change increased the likelihood of identifying an event boundary. Others have similarly found that segmentation is more strongly driven by quantifiable and observable visual features of the videos than it is by knowledge of the activity or its context (Zacks et al., 2009), or by an observer's belief that the activity is goal-directed (Zacks, 2004). The data suggest a prominent role for observable visual features in segmentation, with weak modulatory effects of the observer's internal knowledge or goals. Thus, top-down knowledge and conceptualization of an activity may cause observers to chunk smaller events into larger events, but appear to have little effect on when observers identify boundaries or how those boundaries relate to visual features.

### *1.3. The effects of perspective on observer knowledge and stimulus features*

Most event segmentation research has used videos recorded from the *third-person perspective* (for an exception see Magliano et al., 2014). In these studies, viewpoint is physically separated from the actor and typically shows most, if not all, of the actor's body and location within the broader spatial context. This is consistent with how observers typically view another person's activities. However, events also can be experienced from the actor's own, *first-person perspective*. These can occur with head-mounted cameras, visual imagery, or the spontaneous adoption of an actor's perspective (Tversky & Hard, 2009).

First-person perspectives differ from third-person perspectives in ways that impact both bottom-up sensory input and top-down knowledge and construal of an activity. With first-person perspectives, changes in viewpoint from head or body movements lead to greater variability in visual input, and increase motion and blur in videos. However, objects that are within the actor's reach are viewed up close in first-person perspective videos, making their physical features, identity, and how they might be acted upon more accessible (Borghi, Flumini, Natraj, & Wheaton, 2012; Jackson et al., 2006; Roche &