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
Available online 05 May 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 &

Chainay, 2013; Vogt et al., 2003). In contrast, first-person perspectives provide less information about the actor's size, posture, and location within the larger scene. This could impair the representation of the scene's spatial structure and context (Henderson, Larson, & Zhu, 2008) and the ability to predict the actor's movements through space (Creem-Regehr, Gagnon, Geuss, & Stefanucci, 2013). Therefore, rather than the allocentric (object to object) reference frames afforded by third-person perspectives, first-person perspectives promote the use of egocentric (actor to object) reference frames that are centered on and move with actor (Vogeley & Fink, 2003). If event segmentation depends on these visual features (it is not viewpoint invariant), then viewing an activity from the first-person perspective rather than the third-person perspective should cause an observer to identify different events.

Perspective may also influence how an activity is conceptualized. For example, rotating a key in a lock can be conceptualized as rotating a metal object, starting a car, or heading to work to earn money. These identities vary in how closely they are tied to specific features of the current situation (i.e., object shape and specific muscle movements), the actor's goals, and the actor's character (Vallacher & Wegner, 1987). Importantly, the way an observer conceptualizes an activity may be influenced by whether it was viewed from a first- or third person perspective (Libby & Eibach, 2011). Descriptions of first-person perspectives are more concrete and focused on how an action was performed. Descriptions of third-person perspectives are more abstract, goal-oriented, and focused on their purpose (Libby, Shaeffer, & Eibach, 2009). Similarly, participants who recall an event from a first-person perspective are more likely to attribute behavior to the current situation. Those who recall it from a third-person perspective are more likely to make dispositional attributions (Mcisaac & Eich, 2002; Nigro & Neisser, 1983; Storms, 1973; Taylor & Fiske, 1975). Thus, first- and third-person perspectives promote different ways of conceptualizing an activity. First-person perspectives may lead to more embodied, concrete processing and third-person perspectives may lead to more abstract processing.

1.4. The current study

Event segmentation data suggest a strong and reliable relationship between segmentation and the visual features of an ongoing experience, but offer limited insight into the relationship between segmentation and both the content of the experience and an observer's conceptualization of it. Contrasting first- and third-person perspectives offers a unique way to examine this issue by varying the visual features of an activity while keeping the activity itself constant. First- and third-person perspectives also foreground different aspects of an activity, emphasizing either concrete details of how an activity is performed, in the case of first-person perspective, or abstract conceptualizations of why an activity was performed, in the case of third-person perspective (Libby & Eibach, 2011).

In two experiments we examined whether event segmentation is invariant across first- and third person perspectives. Six activities were simultaneously recorded from a stationary camera (third-person perspective) and from a head-mounted camera (first-person perspective). Participants segmented the videos into events of different grains, and the data were evaluated to determine whether perspective influences how the events were segmented. Though one study previously found that people use information about the current situation to identify events in first-person video games (Magliano et al., 2014), it did not compare segmentation across perspectives. These experiments should provide novel insights into whether segmentation is driven by low-level visual features (similar to viewpoint dependence in object recognition) or if it is also influenced by the content of the video and the observer's conceptualization of the activity (similar to viewpoint invariance). These possibilities predict different effects of perspective on segmentation rate, when events are segmented, and the relationship between segmentation and visual features of the videos.

If segmentation is tied to the low-level visual features of a video such as actor posture and visual change (*visual feature dependent hypothesis*), they should be strongly related to segmentation for both first- and third-person videos. However, segmentation should differ across perspectives to the extent that their visual features differ. Because visual features change more frequently and at different times in first-person videos than in third-person videos, first-person videos should be segmented more frequently and at different times than third-person videos of the same activity. This pattern would be consistent with viewpoint dependence in segmentation.

Alternatively, segmentation may be based on the activity content of the video (*content dependent hypothesis*). If so, observers should segment videos of the same activity in the same way, regardless of whether they view it from a first- or third-person perspective. The relationship between segmentation and the low level visual features of the video, however, should change across perspectives. Thus, event segmentation should be viewpoint invariant in much the same way that object, scene, and action recognition are invariant across luminance changes, orientations, and occlusions (Rust & Stocker, 2010).

A related hypothesis is that the observer's focus on different aspects of the activity will influence segmentation (*focus modulated hypothesis*). First-person perspectives may lead observers to focus on concrete aspects of how an action was performed, while third-person perspectives may lead them to focus more on abstract goals. Because focusing on how an activity is performed increases segmentation rate (Cohen & Ebbesen, 1979), first-person videos may be segmented at a finer grain than third-person videos. In addition, visual features should play a larger role in segmenting first-person videos than third-person videos. Notably, segmentation of first- and third-person videos may still be based on the activity content, but just vary in grain. If so, then boundaries should still occur at similar times in first- and third-person videos.

Experiment 1 tested these hypotheses by examining the effects of perspective on when and how frequently observers segment activities in first- and third-person videos. Experiment 2 was performed as a replication of Experiment 1 and to examine whether perspective influences the organization of fine events within coarse events.

2. Experiment 1

The first experiment examined segmentation of the same activity across different perspectives. Three groups of participants were recruited to segment each activity once at different grains (fine, neutral, and coarse). Each participant segmented half the activities from the first-person perspective, and the other half of the activities from the third-person perspective. If segmentation is viewpoint dependent (visual feature dependent hypothesis), then observers should identify different units of an activity when it is viewed from different perspectives. If segmentation is invariant across perspectives (content dependent hypothesis), then participants who view an activity from different perspectives should identify similar events. However, if differences in the way first- and third-person videos are conceptualized carry over to segmentation (focus dependent hypothesis), then perspective should also influence segmentation rate and the role of visual features in segmentation.

2.1. Methods

2.1.1. Participants

Volunteers were recruited from Cornell University's undergraduate population and the Institutional Review Board approved all methods and procedures. All participants provided informed consent. Participants were compensated $10/hour or with course credit.

Seventy-two participants (51 female; 21 male; age M = 20.19 years, SD = 1.86) completed Experiment 1. Sample size was selected before data collection began following a power analysis with G*Power (Faul,

Erdfelder, Lang, & Buchner, 2007). With $1 - \beta = .9$ and $\alpha = .05$, this experiment had the sensitivity to detect within-between group interactions with effect sizes of $f = 0.214$ (equivalent to $\eta^2_p = .044$) or greater in a 3×2 Analysis of Variance (ANOVA). Data from 5 additional participants were excluded from analyses due to computer errors.

2.1.2. Videos

Actors (4 women, 2 men) were recorded performing six *activities*: eating breakfast (*breakfast,* 253 s long), doing laundry (*laundry,* 231 s long), organizing a desk and bookcase (*office,* 212 s long), making a pasta dinner (*pasta,* 448 s long), building a table (*table,* 361 s long), and clearing toys (*toys,* 317 s long). A different actor performed each activity. Each activity was simultaneously recorded from two vantage points to produce two videos of the same activity from different perspectives (yielding twelve *videos*). For the third-person video a camera (GoPro Hero 4, Silver Edition) was positioned on a stationary tripod. The tripod was positioned as closely as possible to the action while ensuring that the actor and the objects he or she interacted with were visible throughout the video, even as he or she moved around the room. The location was selected to capture as much of the activity from the front or side of the actor as possible. The camera was positioned at about the eye level of a typical adult, 5–6 feet above the floor. Viewing angles were constrained by the layout of the rooms. For the first-person video the actor wore a camera (GoPro Hero 3+, Black Edition) on his or her forehead using an elastic head strap (manufactured by GoPro). The camera was positioned to capture the region of space directly in front of the actor, including the space in which they would act on objects. Adjustments were made prior to recording to ensure that the actor would be able to naturally interact with objects in the camera's view. Fig. 1 shows frames from both perspectives for all 6 activities. The head-mounted camera was visible in the third-person videos. The third-person camera was rarely visible in the first-person videos. All twelve videos were acquired at 1910 × 1080 pixels and 29.98 fps.

Prior to recording the actors rehearsed a rough script to ensure that they interacted with specific objects in the video, but not others, and that they performed some actions before others. For example, in the office activity the actor was asked to dust the bookcase before writing a note. These directions were included for other experiments that will not be reported here. Actors were asked to ensure that their actions would be visible in both videos.

A seventh activity that depicted a man relaxing outside, reading, and using his phone was used as a practice video. This activity was recorded on two separate occasions on the same day and with the same actor. One recording was from the third-person perspective, the other was from the first-person perspective (both used the GoPro Hero 3+, Black Edition). Data from this activity were not analyzed.

2.1.3. Video feature coding

Changes in visual features were calculated for every fifth frame in the videos, after they had been converted to grayscale images. Luminance was defined as the mean pixel value of the frame. Clutter was defined as the proportion of pixels in the frame that were classified as an edge by the Laplacian of Gaussians method in Matlab's edge detection algorithm (Image Processing toolbox). Other measures compared the frame being measured (e.g., frame 5) to a reference frame four places earlier (e.g., frame 1). The *visual activity index* (VAI) captured the amount of pixel by pixel change from a reference frame to the current frame. It was calculated by correlating each pixel value in the current frame and the reference frame and then subtracting the correlation from 1. The VAI was 0 for identical images and 2 for images with

Fig. 1. Frames from the twelve videos, which show six activities from both the first- and third-person perspectives.

inverted contrasts (Cutting, DeLong, & Brunick, 2011; see also Hard et al., 2011; Loucks & Baldwin, 2009 for similar measures). *Optical flow* captured the amount of relative movement in the videos (magnitude squared output of the Lucas-Kanade algorithm in Matlab's Computer Vision System toolbox). The output, which roughly corresponded to the squared spatiotemporal derivative, was then averaged for each frame.

Because the videos were recorded with GoPro cameras, they had fish eye distortions, particularly at their edges. The VAI and optical flow were also measured for versions of the videos that had the distortion removed in Adobe Premiere Pro CC 2016 (which also cropped the videos by 15.5%). The VAI and optical flow mean were strongly correlated across the original and corrected videos (r's = .897 and .752, without regard to perspective). Analyses that used measures from the fisheye corrected videos did not yield substantively different results. Because participants viewed the uncorrected videos the values for these videos were used in the main analyses.

Finally, two features of all 12 videos were coded by two independent raters. *Touch onsets* were coded when either hand visibly contacted an object. For example, an actor picking up a jar with her left hand and then grabbing the lid of the jar with her right hand was two touch onsets. *Touch offsets* were coded when the part of the actor touching the object moved far enough away that the background was visible between them. Because touch onsets and offsets had to be visible, they could differ across perspectives. Features were coded for every video frame, and then binned into 1 s long intervals. With these criteria, raters showed acceptable inter-rater agreement about the beginnings and ends of contact between the actor and an object (Cohen's Kappa = .716). Discrepancies in feature codes binned every second were resolved by the coders through discussion.

2.1.4. Equipment

Segmentation data were acquired on an iMac (2008, OS 10.8.5) with a 20 in. LCD display (1024 × 760 pixel resolution, 60 Hz refresh rate). The tasks were programmed and run in Matlab (Mathworks, Inc) using Psychtoolbox (Brainard, 1997; Pelli, 1997). Testing was performed in an interior, normally lit room. Participants sat approximately 50 cm from the display but were free to move around.

2.1.5. Procedure and design

After participants provided informed consent they completed the segmentation task. For this task videos were presented one at a time in the center of the screen (75% of the horizontal dimension of the screen; aspect ratio was preserved) over a black background. Participants were told they would view videos of people performing everyday activities and to press a button whenever they believed one natural and meaningful unit of activity ended and another began. Breaks were offered between videos.

All participants practiced the segmentation task with the practice video before segmenting half of the activities from the same perspective. Following segmentation of the first group of videos, participants repeated the practice and task for the remaining three activities from the other perspective. The perspective of the practice video matched the perspective of the videos that followed. The assignment of activities to perspective was counterbalanced across participants. In this way, each participant contributed data to both perspectives but viewed each activity once.

Segmentation instructions were varied between participants. For *fine segmentation*, participants identified the smallest (*fine*) units of activity. For *coarse segmentation*, participants identified the largest (*coarse*) units. Finally, participants in the *neutral segmentation* group were not instructed to identify units of a particular size. Participants who received fine or coarse segmentation instructions repeated the practice until they pressed the button within a pre-defined range: 12–30 times (7.3–18.2 times per minute) for fine segmentation and 2–6 times (1.2–3.6 times per minute) for coarse segmentation. Participants were never informed of these ranges but were asked to repeat the practice to identify more (or fewer) activities in the video until performance was within range. Participants in the neutral condition were not required to segment within a preset range, allowing them to freely vary the frequency with which they segmented the activities.

2.1.6. Measures and analyses

Statistical analyses were performed in Matlab 2016b (Mathworks, Inc.) and R (3.3.2), using the core and logistf packages (Heinze & Ploner, 2004) and custom coded routines. Most measures required examining the time series of button presses at the group and individual levels. Each video was divided into 1 s long time periods (*bins*), and a time series indicating whether a button was pressed during a bin was generated for each person who viewed the video (*individual time series*). *Group time series* for each combination of activity, perspective, and segmentation instruction were created by calculating the proportion of participants in that condition who pressed the button within each time bin. Bin sizes were set to 1 s to preserve any variation in the timing of button presses across perspectives. Data are included as Supplementary materials which are available online.

2.2. Results

If segmentation is tied to the low-level visual features of the video, the action content of the video, or the observer's conceptualization, then perspective will produce different effects on three measures of segmentation (Section 1.4): segmentation rate, agreement about the location of event boundaries, and the relationship between segmentation and perceptual features of the activity. Because these predictions assume that visual features differ across perspectives, we first address whether this was the case.

2.2.1. Visual features

For each video we measured luminance, clutter, visual activity (VAI), mean optical flow magnitude, and the onsets and offsets of object touches (Section 2.1.3). Summary statistics are in Table 1. Relative to third-person videos, first-person videos were darker and less cluttered. They also had more visual activity, optical flow, and visible touch onsets. These differences were significant for clutter, VAI, and optical flow, largest $p = .014$, resulted in a trend for luminance, $p = .083$, and were not significant for touch onset or touch offset, smallest $p = .396$. In addition, visual features were weakly to moderately correlated across perspectives (Table 1). Despite capturing the same underlying activity, the visual features of the first- and third-person videos sometimes increased and decreased at different times.

2.2.2. Segmentation rate

Segmentation rates were examined to address two questions. First, did participants follow instructions, segmenting at different rates for fine, coarse, and neutral events? Second, did segmentation rates differ across first- and third-person perspectives, as predicted by the visual feature dependent and focus modulated hypotheses?

To address the first question, a one-way between subjects Analysis of Variance (ANOVA) on the number of button presses per minute indicated significant differences across grains, $F(2, 69) = 23.34$,

Table 1
Mean, standard deviation (in parentheses), and correlations of visual attributes for each activity (N = 6) recorded from first- and third-person perspectives.

	Luminance	Clutter	Flow	VAI	Onset	Offset
First-person	151 (13)	.014 (.001)	.388 (.099)	.206 (.064)	.263 (.094)	.225 (.072)
Third-person	160 (6.9)	.018 (.002)	.026 (.013)	.006 (.004)	.244 (.120)	.231 (.092)
Correlation	.056 (.148)	.140 (.181)	.117 (.135)	.335 (.109)	.568 (.104)	.595 (.107)

Table 2
Means and standard deviations of the number of button presses per minute in each segmentation condition of Experiments 1 and 2.

	Fine	Neutral	Coarse
Experiment 1	10.70 (5.21)	6.01 (5.46)	2.09 (0.87)
Experiment 2	15.59 (6.85)	–	2.85 (1.73)

$p < .001$, $\eta^2_p = .402$. As expected, participants identified the most boundaries under fine segmentation instructions, a moderate number of boundaries under neutral segmentation instructions, and the fewest boundaries under coarse segmentation instructions (Table 2).

Of greater interest was whether participants segmented an activity at different rates when they viewed it from different perspectives. Because the VAI changed more in first-person videos, the visual feature dependent hypothesis predicts that segmentation rates will be greater for first-person videos. The same is true for the focus modulated hypothesis, which predicts that observers will focus more on concrete actions than on abstract goals when segmenting first person-videos. The content dependent hypothesis of segmentation predicts that segmentation rates should be similar across perspectives. To test these predictions, each participant's data were transformed to z-scores to increase statistical power (Bush, Hess, & Wolford, 1993), and then submitted to an ANOVA with perspective and grain as factors (Fig. 2a). The results indicated that standardized segmentation rates were greater for first-person videos than for third-person videos, $F(1, 69) = 5.86$, $p = .018$, $\eta^2_p = .078$. Though the effect of perspective was numerically reversed for the neutral grain, the perspective × grain interaction was not significant, $F(2, 69) = 2.20$, $p = .116$.

2.2.3. Boundary agreement

The hypotheses outlined in Section 1.4 make different predictions about the effect of perspective on boundary identification. The visual feature dependent hypothesis predicts that participants will identify different boundaries in first- and third-person videos of the same activity because their visual features differ. In contrast, the content dependent hypothesis predicts that participants will identify similar boundaries for the same activity, regardless of their perspective. The same is true for the focus modulated hypothesis, which suggests only that the granularity of segmentation will change across first- and third-person videos.

To test these predictions, the degree to which an individual agreed with the boundaries identified by the group of participants who viewed an activity from the *same* perspective and the group who viewed it from the *different* perspective was assessed. These analyses utilized the individual-group agreement measure (Kurby & Zacks, 2011) and rest on the logic that agreement will be greater when individuals and groups identify similar units of activity. Agreement was quantified by calculating the correlation between each individual's time series and the group time series for the same perspective group (excluding the individual's data) and the different perspective group. Correlations were scaled by the minimum and maximum correlations possible, given the number of boundaries the participant identified (as in Kurby & Zacks, 2011).

Individual-group agreement (Fig. 2b) was evaluated in an ANOVA with individual viewer perspective (first vs. third), group perspective (same vs. different), and grain as factors. Consistent with the content dependent and focus modulated hypotheses, individual-group agreement did not significantly differ across same and different perspective groups, $F(1, 69) = 0.14$, $p = .713$, and this factor did not significantly interact with viewer perspective or grain, smallest $p = .214$ for the individual viewer perspective × group perspective × grain interaction. The two other main effects reached or showed a trend toward significance. First, individual-group agreement was numerically greater for individuals viewing the activity from the first-person perspective, resulting in a trend toward a main effect, $F(1, 69) = 3.35$, $p = .072$. This trend implies that viewing an activity from the first-person perspective allows one to more reliably identify boundaries identified from either perspective. Second, grain reliably influenced individual-group agreement, $F(2, 69) = 49.19$, $p < .001$, $\eta^2_p = .588$, reflecting an increase in agreement from coarse to neutral and neutral to fine grained events, all $t(46) > 4.846$, $p < .001$, $d > 1.399$. The agreement data indicate some degree of invariance in segmentation: Observers identified similar boundaries in first- and third-person perspectives.

2.2.4. The relationship between visual features and segmentation

A critical test of the content dependent hypothesis is whether the relationship between visual features and boundary identification changes across perspectives. Because visual features in the first- and third-person videos are only moderately correlated (Table 1), the relationship between visual features and segmentation should change if participants identify boundaries based on activity content. The focus modulated hypothesis suggests that segmentation should be more strongly associated with the visual features of first-person videos than of third person videos because they emphasize concrete rather than abstract information about the activity. Finally, the visual feature based hypothesis suggests boundaries should be associated with changes in visual features to the same degree in both types of videos.

This analysis examined whether the relationship between visual features and segmentation varied across perspectives. Logistic regression models were first fit to each individual's segmentation data. The models included VAI, touch onsets, touch offsets, and their interactions as predictors. VAI was scaled to have a mean of 0 and a standard deviation of 1 within each video. The models utilized the VAI rather than optical flow because it was more strongly correlated across perspectives and, upon visual inspection, appeared to more accurately capture the actors' movements (see Appendix A). Fit was indexed with the penalized log likelihood ratio statistic (PLR).

To evaluate overall model fit, the PLR was averaged across grains

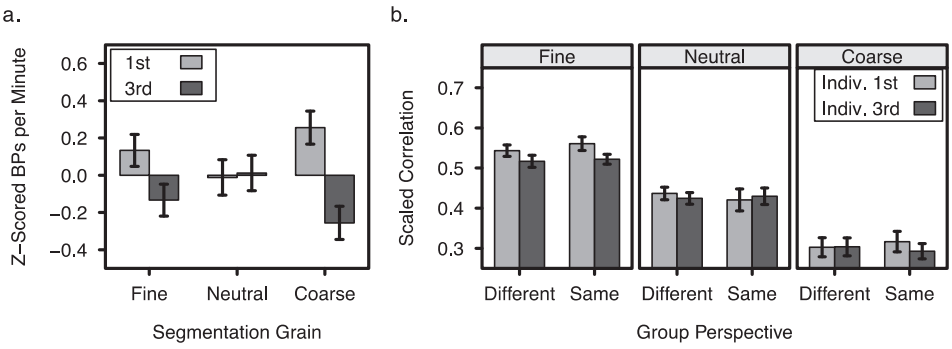

Fig. 2. Segmentation rate (a), indexed by z-scored button presses per minute, and individual-group agreement (b), indexed by the scaled individual-group correlation, in Experiment 1. Error bars indicate ± 1 standard error of the mean.

Table 3
The effects of visual features (VAI, touch onsets, and touch offsets) on the likelihood of a button press and their modulation by perspective and grain in Experiment 1 (N = 72) and Experiment 2 (N = 24).

Exp	Effect	Overall		Perspective			Grain			Persp. × Grain		
		Obs.	*95% CI*	*F*	*p*	η^2_p	*F*	*p*	η^2_p	*F*	*p*	η^2_p
Exp. 1												
	PLR	**18.554**	5.249–8.146	**6.575**	.016	.087	3.057	.057	.081	0.721	.471	.020
	VAI	**0.173**	−0.113 to 0.095	**7.564**	.010	.099	**9.792**	.000	.221	0.802	.432	.023
	Onset	**0.482**	−0.243 to 0.188	0.310	.575	.004	0.208	.817	.006	3.029	.054	.081
	Offset	**0.430**	−0.247 to 0.216	2.552	.103	.036	2.648	.071	.071	0.403	.674	.012
	VAI × Onset	−0.110	−0.183 to 0.236	**4.037**	.026	.055	0.247	.846	.007	0.675	.564	.019
	VAI × Offset	0.036	−0.246 to 0.327	1.898	.183	.027	0.118	.909	.003	0.142	.864	.004
	Onset × Offset	**−0.496**	−0.397 to 0.416	1.091	.280	.016	0.690	.530	.020	2.583	.078	.070
	3-Way	0.031	−0.648 to 0.397	2.529	.090	.035	0.251	.779	.007	1.509	.196	.042
Exp. 2												
	PLR	**16.152**	4.878–7.967	**4.540**	.041	.165	**34.409**	.000	.599	3.400	.075	.129
	VAI	**0.147**	−0.117 to 0.130	**5.337**	.035	.188	**4.806**	.019	.173	1.697	.203	.069
	Onset	**0.484**	−0.283 to 0.259	1.229	.241	.051	2.358	.133	.093	3.423	.073	.130
	Offset	**0.578**	−0.276 to 0.274	0.013	.907	.001	**6.853**	.013	.230	0.102	.752	.004
	VAI × Onset	0.040	−0.292 to 0.312	2.064	.162	.082	2.170	.139	.086	0.017	.886	.001
	VAI × Offset	0.382	−0.486 to 0.524	**8.466**	.014	.269	1.344	.276	.055	0.083	.786	.004
	Onset × Offset	**−0.625**	−0.415 to 0.487	0.034	.850	.001	**8.497**	.009	.270	0.040	.844	.002
	3-Way	−0.423	−0.819 to 0.650	0.899	.358	.038	0.206	.665	.009	0.606	.425	.026

Note: Obs.: Observed value. The observed value for the overall test are averaged across perspectives and grains. 95% CI: interval that captured 95% of the statistics in a simulation of expected values under the null hypothesis. PLR: Penalized Likelihood Ratio for the full model. For Experiment 1, perspective *F* degrees of freedom = 1, 69, grain and perspective × grain interaction *F* degrees of freedom = 2, 69. For Experiment 2, *F* degrees of freedom = 1, 23. Values in bold are unexpected under the null model with $p < .05$.

Table 4
Mean and standard deviation (in parentheses) of the logistic regression coefficients from model fits to individual segmentation data in Experiment 1.

Effect	Fine Grain		Neutral Grain		Coarse Grain	
	First	Third	First	Third	First	Third
VAI	0.140	0.041	0.247	0.150	0.242	0.216
	(0.150)	(0.141)	(0.198)	(0.108)	(0.217)	(0.124)
Onset	0.677	0.329	0.323	0.533	0.523	0.506
	(0.436)	(0.477)	(0.903)	(0.58)	(0.805)	(0.500)
Offset	0.199	0.296	0.567	0.631	0.327	0.562
	(0.699)	(0.399)	(0.712)	(0.445)	(0.819)	(0.632)
VAI × Onset	−0.171	−0.026	−0.261	−0.022	−0.108	−0.069
	(0.236)	(0.34)	(0.775)	(0.214)	(0.336)	(0.231)
VAI × Offset	0.009	0.098	−0.009	0.035	−0.020	0.106
	(0.428)	(0.297)	(0.379)	(0.304)	(0.570)	(0.375)
Onset × Offset	−0.487	−0.22	−0.396	−0.746	−0.411	−0.715
	(0.656)	(0.651)	(1.164)	(0.704)	(1.116)	(0.936)
3-Way	0.032	−0.074	−0.133	0.184	−0.107	0.282
	(0.613)	(0.589)	(1.037)	(0.641)	(0.669)	(0.864)

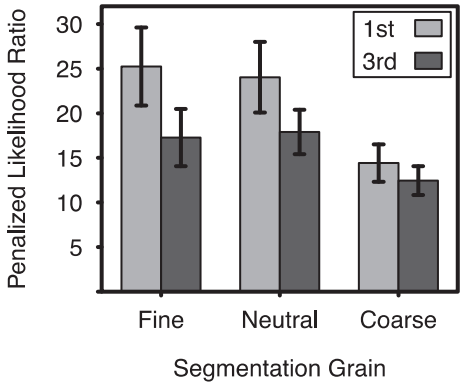

Fig. 3. Penalized likelihood ratios of logistic regression models fit to button presses across perspectives and grains in Experiment 1. Error bars indicate ± 1 standard error of the mean.

and perspectives. The expected value under the null hypothesis was simulated by (1) shuffling the timestamp of the visual features, (2) recalculating the individual models, (3) averaging across grains and perspectives, (4) repeating steps 1–3 1000 times, and (5) constructing 95% confidence intervals around mean simulated values. Visual features were shuffled as a set to maintain covariance in the predictor variables. To evaluate the effects of grain and perspective on model fit, the PLRs from the individual models were evaluated in an ANOVA with grain, perspective, and their interaction as factors. The probability of the observed F statistics under the null hypothesis (no effect of perspective or grain) was estimated using a resampling procedure, which randomized the observed PLR values across participants and conditions before recalculating the F statistics (1000 replications). These procedures were repeated for the observed regression coefficients for the VAI, touch onsets, touch offsets, and their interactions. Statistics from these analyses are reported in Tables 3 and 4.

Consistent with previous work, the regression models describing segmentation as a function of visual features fit the data better than expected by chance (the observed PLRs were greater than those from the simulated null distribution, Table 3). Importantly, these models fit the segmentation data for first-person videos better than they fit the data for third-person videos (main effect of perspective on PLR, Fig. 3). Visual features also provided a poorer fit of the coarse segmentation data than of the fine or neutral segmentation data. This difference showed a trend toward significance (Table 3). Fit was not reliably influenced by the interaction of grain and perspective.

Subsequent analyses of the influence of each visual feature on segmentation, captured by the logistic regression coefficient, suggested that boundaries were more likely to be identified when the VAI increased, when there was a touch onset, and when there was a touch offset (Table 3). Onsets and offsets were under-additive. The weighting of the VAI increased from third- to first-person perspectives (Tables 3 and 4). The observations that visual features, particularly the VAI, better account for segmentation of first-person videos than of third-person videos are consistent with the content dependent and focus modulated segmentation hypotheses, but not the visual feature based hypothesis.

Though the effect of segmentation grain on the relationship between

Table 5
Mean and standard deviation (in parentheses) of the logistic regression coefficients from model fits to individual segmentation data in Experiment 2.

Effect	Fine Grain		Coarse Grain	
	First	Third	First	Third
VAI	0.179 (0.221)	0.020 (0.169)	0.233 (0.369)	0.156 (0.054)
Onset	0.623 (0.590)	0.507 (0.452)	0.209 (0.796)	0.596 (0.678)
Offset	0.753 (0.448)	0.796 (0.461)	0.392 (0.746)	0.373 (0.734)
VAI × Onset	−0.106 (0.422)	0.059 (0.301)	0.032 (0.692)	0.177 (0.313)
VAI × Offset	0.069 (0.418)	0.488 (0.777)	0.318 (0.662)	0.653 (1.025)
Onset × Offset	−0.840 (0.709)	−0.839 (0.622)	−0.451 (1.168)	−0.368 (1.167)
3-Way	−0.271 (0.801)	−0.649 (1.168)	−0.371 (1.080)	−0.401 (1.269)

visual feature and boundary identification is not the focus of this paper, it is worth noting that the effects of perspective were consistent across grains in these analyses. In addition, the relationship between the VAI and segmentation was smaller for fine boundaries than for neutral or coarse boundaries (Table 5 and Fig. 5). This difference is consistent with previous reports indicating that coarse boundaries are associated with larger frame to frame changes in visual content than are fine boundaries (Hard et al., 2011).

2.3. Discussion

Experiment 1 provided initial evidence in favor of both the content dependent and focus modulated hypotheses: The events participants identified in first- and third-person videos could not be distinguished from each other, reflecting a change in the relationship between boundaries and the video's visual features. However, consistent with the hypothesis that focus affects segmentation, participants segmented at a higher rate when they viewed first-person videos rather than third-person videos during fine segmentation tasks. Experiment 2 was performed to determine whether these effects replicate in a second group of participants.

3. Experiment 2

The data from Experiment 1 are inconsistent with the visual feature dependent hypothesis. They also suggest that segmentation may be based on more concrete conceptualizations of an activity when it is viewed from the first-person perspective rather than a third-person perspective. Experiment 2 was performed to test whether these effects replicate in a new sample of participants.

Another goal of Experiment 2 was to explore the effect of perspective on the hierarchical organization of fine events within coarse events. Differences in the availability of particular visual features (e.g., actor posture, large visual changes from head motion) in first-person videos could make it difficult to group fine events into coarse events. For example, changes in the causal structure of events may be used to identify coarse boundaries (e.g., Zacks et al., 2010), and these changes may be less visible in first person videos that offer a limited field of view. Focusing on concrete, rather than abstract information about an activity in first-person videos could have a similar effect. Therefore, one implication of the focus modulated hypothesis could be that third-person videos may be segmented more hierarchically than third-person videos. In contrast, if segmentation is based on activity content, then perspective should have little effect on the hierarchical organization of events. To test this possibility, each participant in Experiment 2 segmented each video into fine and coarse events and measures of hierarchical segmentation were examined.

3.1. Methods

3.1.1. Participants

A new group of 24 participants completed Experiment 2 (13 female; 11 male; age M = 20.13, SD = 1.28). For within-group paired samples t-tests, $1 - \beta = .9$, and $\alpha = .05$, a sample size of 24 was sensitive to effect sizes of $d = 0.69$ or greater.

3.1.2. Equipment

Data were acquired on a Dell 7010 Windows 7 PC with a 17 in. CRT display (1024 × 760 pixel resolution, 75 Hz refresh rate) using Matlab.

3.1.3. Procedure and design

Participants in Experiment 2 segmented two activities (office and laundry, selected because they were shortest in length) into fine and coarse events from both the first and the third-person perspectives. Videos and tasks were ordered so activity changed most frequently, followed by grain, and finally perspective. Thus, participants segmented each activity at both grains before perspective changed. Activity order, grain order, and perspective order were counterbalanced across participants.

3.2. Results

To replicate the findings from Experiment 1, the data from Experiment 2 should show the following: (1) participants identified more boundaries in first-person videos than in third-person videos, (2) individual-group agreement on boundary location is similar when the individual viewed the activity from either the same or different perspective as they comparison group, and (3) the relationship between visual features and boundary identification is greater for first-person videos than for third-person videos. These findings argue against the visual feature dependent hypothesis, which predicts that differences in visual features in first- and third-person videos should cause them to be segmented differently (contrary to finding number 2) and that the relationship between visual features and segmentation should be consistent across video types (contrary to finding 3).

3.2.1. Segmentation rate

Participants in Experiment 2 followed instructions and pressed the button more often when they were instructed to identify fine events (Table 1), $t(23) = 10.26$, $p < .001$, $d = 2.55$. In addition, standardized segmentation rates were numerically greater for first-person videos than third-person videos (Fig. 3a), though this resulted in a trend toward a significant main effect, $F(1, 23) = 3.33$, $p = .081$. The main effect of grain and the perspective × grain interaction were not significant, $F(1, 23) = 0.10$, $p = .753$ and $F(1, 23) = 0.59$, $p = .449$, respectively. The segmentation rate data therefore followed the same pattern as Experiment 1, though the perspective effect was not as clearly present.

3.2.2. Boundary agreement

The feature dependent hypothesis predicts that participants will segment first- and third-person movies at different times. To evaluate this possibility, individual-group agreement was analyzed with a repeated measures ANOVA that included viewer perspective, group perspective, and grain as factors. Unlike Experiment 1, agreement was greater when individuals and groups segmented an activity from the same perspective, rather than the different perspective, resulting in a main effect of group perspective, $F(1, 23) = 6.74$, $p = .016$, $\eta^2_p = .226$. This effect was qualified by a significant three-way interaction between viewer perspective, group perspective, and grain, $F(1, 23) = 4.67$, $p = .041$, $\eta^2_p = .169$. Overall, agreement was greater during fine segmentation, $F(1, 23) = 105.7$, $p < .001$, $\eta^2_p = .821$. There were no other significant main effects or interactions, smallest $p = .193$, for the group perspective × grain interaction.

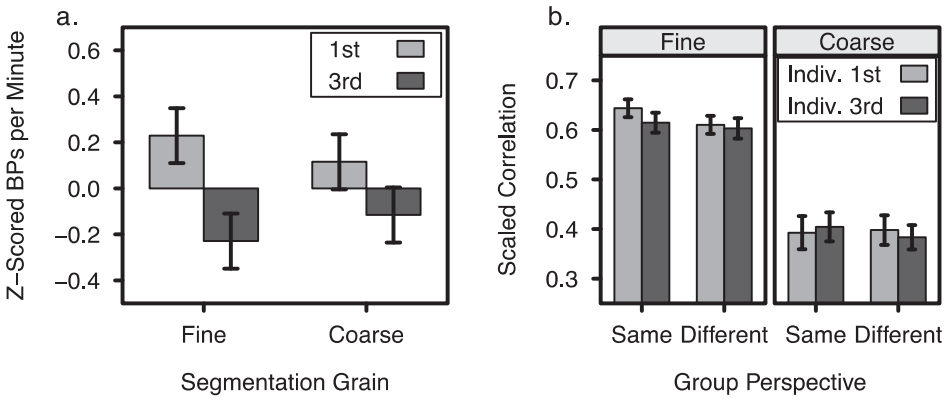

Fig. 4. Segmentation rate (a), indexed by z-scored button presses per minute and individual-group agreement (b), indexed by the scaled individual-group correlation, in Experiment 2. Error bars indicate ± 1 standard error of the mean.

An examination of Fig. 4b suggests that the three-way interaction was driven by higher levels of fine boundary agreement between first-person viewers and the same perspective group (furthest left bar), but comparable agreement among all other fine boundary conditions. Consistent with this characterization, post hoc comparisons indicated that same and different group agreement differed significantly only when first-person viewers identified fine events ($p < .01$ to correct for multiple comparisons): paired t-tests for fine segmentation of first-person viewers, $t(23) = 4.32$, $p < .001$, $d = 0.38$, and of third-person viewers, $t(23) = 1.78$, $p = .088$; paired t-tests for coarse segmentation of first-person viewers, $t(23) = -0.29$, $p = .775$, and of third-person viewers, $t(23) = 1.85$, $p = .077$. In addition, first-person viewers agreed with third-person group boundaries (different group) about as much as did third-person viewers (same group), $t(23) = -0.34$, $p = .736$. This pattern makes straightforward conclusions about whether individuals identified different fine boundaries from different perspectives difficult. If perspective influences the identification of fine boundaries, then same group agreement should be greater than different group agreement for both first- and third-person viewers, not just first-person viewers. In addition, third-person viewers should agree more with third-person group boundaries than first-person viewers. Neither of these patterns was present.

3.2.3. The relationship between visual features and segmentation

The data from Experiment 1 indicated that visual features were associated with segmentation, but that this relationship was stronger for first-person videos than for third-person videos. Experiment 2 replicated this pattern, and argues against the visual feature dependent hypothesis. Model fit (PLR) was reliably better for first-person videos than for third-person videos, and for fine boundaries than for coarse boundaries (Fig. 5). Similarly, in Experiment 2 button presses were more likely when the VAI increased, there was a touch onset, or there was a touch offset (Table 3). The effects of touch onsets and offsets were again under-additive. As before, the effect of VAI on segmentation was greater for first-person videos and for coarse rather than fine events (Table 5). The data show that VAI, touch onsets, touch offsets, and the onset × offset interaction consistently affected segmentation across experiments. The relationship between VAI and segmentation was consistently stronger for first-person videos than for third-person videos, supporting the focus modulated segmentation hypothesis.

3.2.4. The effect of perspective on alignment and enclosure

Changing or removing the accessibility of some visual features in a video could reduce the ability to organize fine events into coarse events. In addition, if observers conceptualize activities more abstractly when they are viewed from the third-person perspective rather than the first-person perspective, participants may better organize them into hierarchical, part-whole structures. The content based hypothesis predicts hierarchical organization should be the same in both perspectives.

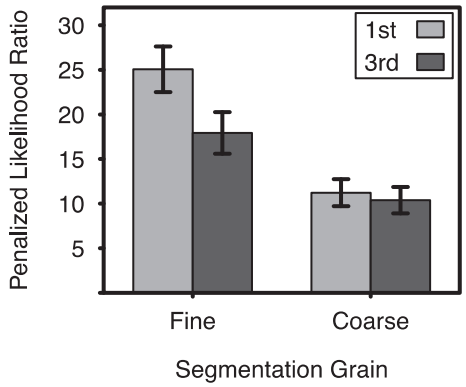

Fig. 5. Penalized likelihood ratios of logistic regression models fit to button presses across perspectives and grains in Experiment 2. Error bars indicate ± 1 standard error of the mean.

The hierarchical organization of fine events into coarse events is indexed by the *alignment* of coarse and fine boundaries. If coarse boundaries are aligned with the nearest fine boundary then the *observed* distance between them should be less than the distance *expected* by chance (Zacks et al., 2001). Hierarchical organization is also evident in the tendency for coarse boundaries to follow, rather than precede the nearest fine boundary (*enclosure*; Hard et al., 2006). There was no evidence that either measure differed across first- and third-person videos (Table 6). An ANOVA indicated that alignment was better than expected by chance, $F(1, 23) = 52.75$, $p < .001$, $\eta^2_p = .696$, but did not vary across perspectives, main effect of perspective, $F(1, 23) = 0.22$, $p = .644$, perspective × score type interaction, $F(1, 23) = 0.49$, $p = .489$. In addition, a *t*-test indicated that enclosure was greater than chance (.5), $t(23) = 5.05$, $p < .001$, $d = 1.03$ (averaged across perspectives), but did not differ across perspectives, $t(23) = -0.14$, $p = .886$. Participants grouped fine events within coarse events similarly across perspectives, again suggesting that perspective has minimal effects on event segmentation (consistent with the content based hypothesis).

Table 6
Mean observed and expected alignment and enclosure scores for each perspective in Experiment 2, with standard deviations in parentheses.

	Alignment (seconds)		Enclosure	
	First	Third	First	Third
Observed	1.60 (1.82)	1.59 (1.47)	.614 (.166)	.620 (.141)
Expected	2.65 (1.74)	2.86 (1.94)	.5	.5

3.2.5. Discussion

Three main findings resulted from Experiment 1: (1) greater segmentation rates of first-person than third-person videos, (2) no difference in individual-group agreement across same and different groups, and (3) stronger association between visual features and segmentation of first-person videos than third-person videos. Only the third finding was unequivocally replicated in Experiment 2. Though segmentation rates were numerically greater for first-person videos, the difference resulted in a trend toward a main effect of perspective. Individual-group agreement was similar across same and different groups in most conditions, but a three-way interaction indicated that perspective may have a small effect on fine boundary identification. It is possible that additional experimental power would have revealed more consistent effects of perspective on boundary identification. However, in Experiment 2 agreement was comparable to levels reported in studies that exclusively used third-person videos (.60–.64 in Experiment 2 and .61–.69 in Kurby & Zacks, 2011). In addition, the effect of perspective was less than half the size of the differences observed across age groups (Kurby & Zacks, 2011). This, combined with a non-replication of this difference in Experiment 1, suggest that the effect of perspective on individual-group agreement is small if it is reliably present.

The strongest and most consistent finding across the two experiments is the observation that visual features are more strongly related to segmentation of first-person videos than third-person videos. Experiment 2 also evaluated the effect of perspective on the hierarchical organization of fine into coarse events. A lack of an effect of perspective in this measure further affirms that segmentation is consistent across perspectives. This finding is consistent with the content based hypothesis.

4. General discussion

Changing the perspective from which an activity is viewed changes the visual features that an observer can use to segment it. First- and third-person videos in this study differed in clutter, visual activity, and in whether touch onsets or offsets were visible. They also differed in which features were accessible and visible (e.g., actor posture, hands, body, and head). It is not surprising that first-person videos changed more than third-person videos, or that visual features were not strongly correlated across perspectives. What is remarkable, however, is that these visual differences had small and inconsistent effects on the way participants divided the activity into parts. By varying the features of an activity without changing the activity itself, Experiments 1 and 2 revealed that segmentation is relatively robust to changes in the visual input, ruling out the visual feature dependent hypothesis. Instead, segmentation mechanisms appear to flexibly use visual information to identify the structure of the underlying activity in a manner that is mostly viewpoint invariant.

4.1. Invariance in segmentation and action recognition

Recent findings have emphasized the central importance of visual features in segmentation. These include the findings that people identify similar boundaries when the video is played backward (Hard et al., 2006), that similar boundaries are identified for visually rich videos and visually sparse videos with similar motion characteristics (Zacks et al., 2009), that a large proportion of variance in segmentation behavior can be explained by variance in visual features (Cutting et al., 2012), and that eye movements are affected more by visual features and editing techniques than by observer knowledge (Loschky, Larson, Magliano, & Smith, 2015). In all of these studies, visual information may be important because it is directly used by segmentation mechanisms or it may be important because it is correlated with other information that is (e.g., knowledge about action, intentionality, etc.).

Experiments 1 and 2 suggest that the relationship between visual features and segmentation exists because those features are correlated with higher-level visual or conceptual features of the activity itself. Despite clear and repeated evidence that body posture is correlated with segmentation in prior work (Hard et al., 2006; Newtson et al., 1977), observers in Experiments 1 and 2 identified similar events when they could see body posture (in third-person videos) and when they could not (in first-person videos). Therefore, being able to directly observe an actor's body and posture is not necessary for segmentation. This finding mirrors robust action recognition when motion trajectories vary (Loucks & Baldwin, 2009) as well as findings that visual transients and discontinuities produced by cuts in narrative film do not lead to segmentation on their own (Baker & Levin, 2015; Magliano & Zacks, 2011).

These data provide initial evidence of invariance in segmentation behavior, but cannot speak directly to how it might be achieved. This is particularly true given the status of the invariance debate in object recognition (Biederman, 1987; Edelman & Bülthoff, 1992; Gauthier & Tarr, 2016). Recent approaches suggest that, to the degree it exists, invariance in object recognition arises from the multi-dimensional and hierarchical structure of visual processing and may be task or context dependent (DiCarlo & Cox, 2007; Gauthier & Tarr, 2016; Rust & Stocker, 2010). Considerations of how actions are represented and influenced by task and context are likely to provide the most insight into invariance in segmentation. It is possible that viewpoint dependence will be more evident early in development, when events are novel, or if viewpoint influences the ability to distinguish the object types, tokens, and states that are part of an event (Hindy, Solomon, Altmann, & Thompson-Schill, 2015). Explorations of when invariance emerges in development and learning will be highly informative.

Two additional questions along these lines should be addressed in future research. First, understanding of the effects of perspective in segmentation will benefit from examining the role of high-level conceptual features of an activity to the segmentation of first- and third-person videos. For example, research that has examined segmentation of narrative text, film, and picture stories (e.g., Magliano, Kopp, McNerney, Radvansky, & Zacks, 2012; Magliano et al., 2014; Zacks et al., 2009, 2010), have demonstrated that changes in actor goals, object interactions, and actions are associated with boundary identification in segmentation tasks. If these features can be reliably coded and identified within first- and third-person videos (see Yordanova et al., 2017, for why this may not be trivial), then high-level conceptual features of an activity could be used to segment it from different perspectives.

Second, it will be important to evaluate whether the information used to identify actions in machine vision is present in both first-person and third-person videos, and, if it is, whether people use similar types of information to segment events. One common approach to action recognition in machine vision is to identify spatiotemporal interest points (luminance changes in space and time) and map these to action templates (e.g., Peng et al., 2016). Other approaches track hand or head motion (e.g., Singh, Arora, & Jawahar, 2016) or object locations and states over time (e.g., Lea, Reiter, Vidal, & Hager, 2016) in first- or third-person videos. Both types of visual information may be more diagnostic of an event boundary than the VAI. An application of these algorithms to the videos used in this study could therefore provide additional insight into the types of information people use to segment events.

4.2. The role of perspective in segmentation

Perspective influences the way people construe and represent the activities of others and of themselves. The data from this study suggest this may affect the relationship between segmentation and visual features. Beyond differences in their visual features, first- and third-person perspectives play different roles in narrative comprehension, episodic memory, the integration of new knowledge into one's sense of self, and attributions (Borghi, Glenberg, & Kaschak, 2004; Brunyé, Ditman,

Mahoney, Augustyn, & Taylor, 2009; Libby & Eibach, 2011; Mcisaac & Eich, 2002; Nigro & Neisser, 1983; Storms, 1973; Taylor & Fiske, 1975). In addition, first-person perspectives facilitate an observer's access to motor routines that are employed when performing an activity (Jackson et al., 2006; Vogt et al., 2003). Access to this type of information may be important for segmentation: The ability to perform an activity is related to the way an observer segments and represents it in memory (Bailey et al., 2013).

Though less obvious, perspective and sensory information may play a role in the comprehension of events that are communicated through language as well as events that are viewed. Narrated events appear to be segmented in a manner that is similar to observed events (Magliano et al., 2001; Speer, Zacks, & Reynolds, 2007), and may lead readers and listeners to represent their physical, motoric, and emotional features (Borghi et al., 2004; Ruby & Decety, 2001; Speer, Reynolds, Swallow, & Zacks, 2009; Stanfield & Zwaan, 2001). There is some evidence that perspective in narrative influences the way listeners represent the events they describe (Abelson, 1975; Franklin & Tversky, 1990; Franklin, Tversky, & Coon, 1992). For example, mental representations of narrated events appear to be richer and longer lasting when pronouns that encourage embodiment ("you" rather than "I") are used (Brunyé, Ditman, Mahoney, & Taylor, 2011; Brunyé et al., 2009; Ditman, Brunyé, Mahoney, & Taylor, 2010; Ruby & Decety, 2001). Despite the evidence that perspective influences event understanding, however, it did not consistently influence the way participants divided an activity into meaningful events. The functional properties of first- and third-person perspectives in other domains had small effects on event segmentation.

Perspective did, however, change the relationship between segmentation and the visual features of the videos. This pattern is consistent with views that suggest that first- and third-person perspectives differ along a concrete-abstract dimension (Libby & Eibach, 2011). If this were the case, then segmentation of first-person videos would be more closely tied to perceptual information than segmentation of third-person videos. This is exactly what was observed in Experiments 1 and 2. In addition, if it is reliable, the tendency to identify smaller events for first-person videos in Experiment 1 (and, less clearly in Experiment 2) may similarly reflect a greater emphasis on how the activity is performed, rather than on the actor's goals. Whereas verbal descriptions of fine events tend to emphasize actions on objects, descriptions of coarse events tend to emphasize goals (Cohen & Ebbesen, 1979; Zacks et al., 2001).

The present data bear only on the question of whether perspective influenced segmentation, and do not speak to whether perspective influenced other aspects of event encoding and memory. For example, when perspective is manipulated observers' descriptions of an activity may differ in their focus on how (sensorimotor features, emotional states, etc.) or why (goals, actor traits, and consequences) the activity was performed, even when grain is held constant. In addition, presenting an activity from different perspectives may alter the way it is remembered, possibly encouraging the representation of more perceptual detail from first-person videos. Additional research is needed to explore these possibilities.

4.3. Mechanisms of event segmentation

The finding that segmentation behavior is similar across perspectives constrains theory on event segmentation. Currently, theories of event segmentation focus on three types of mechanisms: model-based prediction (EST; Zacks et al., 2007), statistical learning (Avrahami & Kareev, 1994; Buchsbaum et al., 2015; Endress & Wood, 2011), and post hoc comparison (Baker & Levin, 2015; Hymel, Levin, & Baker, 2015). The data from Experiments 1 and 2 suggest that adult observers are unlikely to segment familiar activities based purely on viewpoint dependent information. We explore the implication of this finding for each type of approach below.

According to EST (Zacks et al., 2007), observers use sensory input in conjunction with knowledge of event types (schemata) to build an internal model of the current situation (*event models*). These models generate predictions about upcoming perceptual input that are compared to the actual input. If prediction error is too high, the model is reset and updated to reflect the new situation. Because prediction error signals may be distributed throughout the brain (Clark, 2016; Friston, 2009; Henson & Gagnepain, 2010; Summerfield & Egner, 2009), mismatches in prediction and input from low-level sensory processing to high-level context representations, could be important for determining the adequacy of an event model. One implication of Experiments 1 and 2 is that low-level sensory predictions play a minor role in segmentation. Frame to frame change in visual input was correlated with when participants segmented both types of videos, but differences in VAI did not cause participants to identify different events in most cases. Thus, within the EST framework, error monitoring and gating mechanisms are likely to be sensitive high-level representations of information that is abstracted or inferred from the visual features of the videos. This may include parts of action, biological motion patterns, object states, or the spatial relationship between actors and objects, which observers can identify across varying perspectives (Endress & Wood, 2011; Epstein, Higgins, & Thompson-Schill, 2005; Grossman, Jardine, & Pyles, 2010).

Other accounts of segmentation also emphasize sequential structure in events, which can be used to predict sensory input and chunk experience into units (Avrahami & Kareev, 1994). A variety of statistical learning mechanisms have been explored, including transitional probabilities (which capture the ability to predict the next item in a sequence, e.g., Baldwin et al., 2008), position based learning (which captures when actions occur relative to other salient events, e.g., Endress & Wood, 2011), and temporal community structure (which captures clustering of items in time, Schapiro et al., 2013). This statistical information may be learned and used to mark boundaries between units or chunks of items. The data from Experiments 1 and 2 suggest that once a chunk is learned, it must be recognizable from a variety of perspectives and when information is missing. Position based statistical learning mechanisms can support chunk recognition across multiple third-person viewpoints (Endress & Wood, 2011) and therefore may play a larger role in segmentation than mechanisms that do not generalize to new perspectives.

Finally, at least one account of segmentation suggests that it occurs via post hoc comparison processes, which are triggered by changes in spatial layouts (Baker & Levin, 2015; Hymel et al., 2015). Baker and Levin (2015) suggest that although representations of the external world are limited (e.g., Hayhoe & Ballard, 2005), the spatial configuration of a recently encountered scene is maintained in memory and compared to current perceptual input. Changes in spatial configurations lead to segmentation. The data from Experiments 1 and 2 may be consistent with this claim. Allocentric representations of spatial configurations by the hippocampus may be used to generate egocentric representations of spatial configurations (Bird & Burgess, 2008). In addition, spatial configurations of scenes and events are recognizable from multiple third-person perspectives, particularly when they carry semantic information (Epstein et al., 2005; Huff, Schwan, & Garsoffky, 2011). However, first-person videos provided neither direct access to the complete spatial layout of the scene nor the viewpoint stability of third-person videos. As a result, representations of the scene will need to be built up over much longer periods of time and from restricted field of views that rapidly change.

Though there is more work to do be done, a sketch of the types of information that are used to break experience into meaningful events has begun to emerge. Rather than being tied directly to specific low-level visual features, segmentation is likely based on information extracted from sensory input that varies in proximity, orientation, the presence or absence of some types of visual information, and even scene layouts across different perspectives. Spatial configurations and sequential structure are also important. Together, these findings point to

representations that associate high-level representations of items with their spatiotemporal context. This type of representation could be realized by the hippocampus and its connections to cortex (Bird & Burgess, 2008; Davachi, 2006; Eichenbaum, 2004; Smith & Mizumori, 2006). Indeed, a growing literature explicitly links event segmentation to hippocampal function (Bailey et al., 2013; DuBrow & Davachi, 2013; Ezzyat & Davachi, 2010; Swallow et al., 2011). Invariance in segmentation supports this relationship.

4.4. Conclusion

Since a paradigm to study it was first developed, segmentation has been linked to visual features of experiences, such as actor posture, object motion, and changes in spatial location (Kurby & Zacks, 2008). In contrast, two experiments demonstrated that differences in the visual features of first- and third-person perspectives had little effect on event segmentation. Though real and reliable, the relationship between segmentation and visual features appears to be mediated by content.

Appendix A

See Fig. A1.

Acknowledgments

Portions of this research were presented at annual meetings of the Society for Cognitive Studies of the Moving Image (2016), Cognitive Science Society (2016), and the Psychonomic Society (2016). The authors would like to thank James Cutting for providing code to analyze video features, Jeff Zacks and Thomas Serre for suggestions regarding data analysis, Melissa Elston, Catalina Iricinschi, Erin Isbilen, Vivek Iyer, Gina Mason, and Kedarnath Vilankar for their help with stimulus creation, and Maamie Asamoah-Mensah, Kimberly Lee, Logan Lin, Alyssa Phelps, and Hyung Sop Shin and for their help with data collection and stimulus coding.

Funding

Jovan Kemp was supported by Ronald E. McNair Post-baccalaureate Achievement Program. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

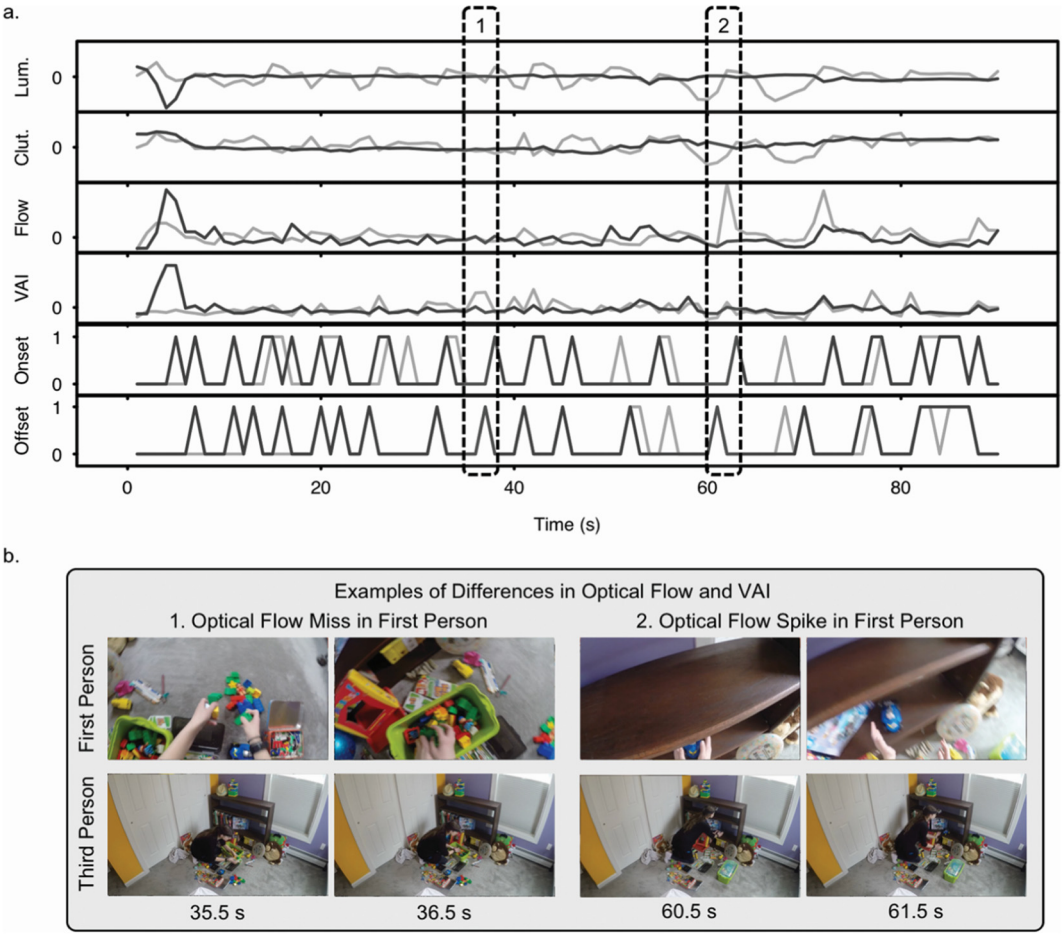

Fig. A1. Visual features of the first 90 s of the toy activity from the first- (gray lines) and third- (black lines) person perspectives (a). Values for luminance, clutter, optical flow, and VAI have been standardized (M = 0; SD = 1). Time periods 1 and 2, indicated by the vertical rectangles, are examples of optical flow diverging for two similar shifts in the actor's body position in the first-person video. In time period 1, optical flow is low despite movement in the actor's head and upper body. In time period 2, optical flow spikes for a similar shift in the actor's position. VAI increased in both instances. Frames from the third-person video show the actor's position changed in both instances.

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2018.04.019.

References

Abelson, R. P. (1975). Does a story understander need a point of view? In *Proceedings of the 1975 workshop on theoretical issues in natural language processing* (pp. 140–143). Stroudsburg, PA, USA: Association for Computational Linguistics. http://dx.doi.org/10.3115/980190.980230.

Avrahami, J., & Kareev, Y. (1994). The emergence of events. *Cognition, 53*, 239–261.

Bailey, H. R., Kurby, C. A., Giovannetti, T., & Zacks, J. M. (2013). Action perception predicts action performance. *Neuropsychologia, 51*(11), 2294–2304. http://dx.doi.org/10.1016/j.neuropsychologia.2013.06.022.

Baird, J. A., & Baldwin, D. A. (2001). Making sense of human behavior: Action parsing and intentional inference. In B. F. Malle, & L. J. Moses (Eds.). *Intentions and intentionality: Foundations of social cognition* (pp. 193–206). Cambridge, MA: MIT Press.

Baker, L. J., & Levin, D. T. (2015). The role of relational triggers in event perception. *Cognition, 136*, 14–29. http://dx.doi.org/10.1016/j.cognition.2014.11.030.

Baldwin, D. A., Andersson, A., Saffran, J. R., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition, 106*, 1382–1407.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*(2), 114–147.

Bird, C. M., & Burgess, N. (2008). The hippocampus and memory: Insights from spatial processing. *Nature Reviews Neuroscience, 9*(3), 182–194.

Blakemore, S.-J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience, 2*(8), 561–567. http://dx.doi.org/10.1038/35086023.

Borghi, A. M., Flumini, A., Natraj, N., & Wheaton, L. A. (2012). One hand, two objects: Emergence of affordance in contexts. *Brain and Cognition, 80*(1), 64–73. http://dx.doi.org/10.1016/j.bandc.2012.04.007.

Borghi, A. M., Glenberg, A. M., & Kaschak, M. P. (2004). Putting words in perspective. *Memory & Cognition, 32*(6), 863–873. http://dx.doi.org/10.3758/BF03196865.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Brunyé, T. T., Ditman, T., Mahoney, C. R., Augustyn, J. S., & Taylor, H. A. (2009). When "you" and "I" share perspectives: Pronouns modulate perspective taking during narrative comprehension. *Psychological Science, 20*(1), 27–32. http://dx.doi.org/10.1111/j.1467-9280.2008.02249.x.

Brunyé, T. T., Ditman, T., Mahoney, C. R., & Taylor, H. A. (2011). Better you than I: Perspectives and emotion simulation during narrative comprehension. *Journal of Cognitive Psychology, 23*(5), 659–666. http://dx.doi.org/10.1080/20445911.2011.559160.

Buchsbaum, D., Griffiths, T. L., Plunkett, D., Gopnik, A., & Baldwin, D. (2015). Inferring action structure and causal relationships in continuous sequences of human action. *Cognitive Psychology, 76*, 30–77. http://dx.doi.org/10.1016/j.cogpsych.2014.10.001.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences, 11*(2), 49–57. http://dx.doi.org/10.1016/j.tics.2006.11.004.

Bush, L. K., Hess, U., & Wolford, G. (1993). Transformations for within-subject designs: A Monte Carlo investigation. *Psychological Bulletin, 113*(3), 566–579. http://dx.doi.org/10.1037/0033-2909.113.3.566.

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* Oxford University Press.

Cohen, C. E., & Ebbesen, E. B. (1979). Observational goals and schema activation: A theoretical framework for behavior perception. *Journal of Experimental Social Psychology, 15*, 305–329.

Creem-Regehr, S. H., Gagnon, K. T., Geuss, M. N., & Stefanucci, J. K. (2013). Relating spatial perspective taking to the perception of other's affordances: Providing a foundation for predicting the future behavior of others. *Frontiers in Human Neuroscience, 7*, 596. http://dx.doi.org/10.3389/fnhum.2013.00596.

Cutting, J. E. (2014). How light and motion bathe the silver screen. *Psychology of Aesthetics, Creativity, and the Arts, 8*(3), 340–353. http://dx.doi.org/10.1037/a0036174.

Cutting, J. E., Brunick, K. L., & Candan, A. (2012). Perceiving event dynamics and parsing Hollywood films. *Journal of Experimental Psychology: Human Perception and Performance, 38*(6), 1476–1490. http://dx.doi.org/10.1037/a0027737.

Cutting, J. E., DeLong, J. E., & Brunick, K. L. (2011). Visual activity in Hollywood film: 1935 to 2005 and beyond. *Psychology of Aesthetics, Creativity, and the Arts, 5*(2), 115–125. http://dx.doi.org/10.1037/a0020995.

Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in Neurobiology, 16*, 693–700.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences, 11*(8), 333–341. http://dx.doi.org/10.1016/j.tics.2007.06.010.

Ditman, T., Brunyé, T. T., Mahoney, C. R., & Taylor, H. A. (2010). Simulating an enactment effect: Pronouns guide action simulation during narrative comprehension. *Cognition, 115*(1), 172–178. http://dx.doi.org/10.1016/j.cognition.2009.10.014.

DuBrow, S., & Davachi, L. (2013). The influence of context boundaries on memory for the sequential order of events. *Journal of Experimental Psychology: General, 142*(4), 1277–1286. http://dx.doi.org/10.1037/a0034024.

DuBrow, S., & Davachi, L. (2014). Temporal memory is shaped by encoding stability and intervening item reactivation. *The Journal of Neuroscience, 34*(42), 13998–14005. http://dx.doi.org/10.1523/JNEUROSCI.2535-14.2014.

Edelman, S., & Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research, 32*(12), 2385–2400. http://dx.doi.org/10.1016/0042-6989(92)90102-O.

Eichenbaum, H. (2004). Hippocampus: Cognitive processes and neural representations that underlie declarative memory. *Neuron, 44*(1), 109–120. http://dx.doi.org/10.1016/j.neuron.2004.08.028.

Endress, A. D., & Wood, J. N. (2011). From movements to actions: Two mechanisms for learning action sequences. *Cognitive Psychology, 63*(3), 141–171. http://dx.doi.org/10.1016/j.cogpsych.2011.07.001.

Epstein, R. A., Higgins, J. S., & Thompson-Schill, S. L. (2005). Learning places from views: Variation in scene processing as a function of experience and navigational ability. *Journal of Cognitive Neuroscience, 17*(1), 73–83. http://dx.doi.org/10.1162/0898929052879987.

Ezzyat, Y., & Davachi, L. (2010). What constitutes an episode in episodic memory? *Psychological Science*. http://dx.doi.org/10.1177/0956797610393742.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: From action organization to intention understanding. *Science, 308*(5722), 662–667. http://dx.doi.org/10.1126/science.1106138.

Franklin, N., & Tversky, B. (1990). Searching imagined environments. *Journal of Experimental Psychology: General, 119*(1), 63–76. http://dx.doi.org/10.1037/0096-3445.119.1.63.

Franklin, N., Tversky, B., & Coon, V. (1992). Switching points of view in spatial mental models. *Memory & Cognition, 20*(5), 507–518. http://dx.doi.org/10.3758/BF03199583.

Friend, M., & Pace, A. (2011). Beyond event segmentation: Spatial- and social-cognitive processes in verb-to-action mapping. *Developmental Psychology, 47*(3), 867–876. http://dx.doi.org/10.1037/a0021107.

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13*(7), 293–301. http://dx.doi.org/10.1016/j.tics.2009.04.005.

Gauthier, I., & Tarr, M. J. (2016). Visual object recognition: Do we (finally) know more now than we did? *Annual Review of Vision Science, 2*(1), 377–396. http://dx.doi.org/10.1146/annurev-vision-111815-114621.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin and Review, 9*(3), 558–565.

Graziano, W. G., Moore, J. S., & Collins, J. E. (1988). Social cognition as segmentation of the stream of behavior. *Developmental Psychology, 24*(4), 568–573. http://dx.doi.org/10.1037/0012-1649.24.4.568.

Grossman, E. D., Jardine, N. L., & Pyles, J. A. (2010). FMR-adaptation reveals invariant coding of biological motion on human STS. *Frontiers in Human Neuroscience, 4*. http://dx.doi.org/10.3389/neuro.09.015.2010.

Hard, B. M., Recchia, G., & Tversky, B. (2011). The shape of action. *Journal of Experimental Psychology: General, 140*(4), 586–604. http://dx.doi.org/10.1037/a0024310.

Hard, B. M., Tversky, B., & Lang, D. S. (2006). Making sense of abstract events: Building event schemas. *Memory & Cognition, 34*(6), 1221–1235.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4), 188–194.

Heinze, G., & Ploner, M. (2004). *A SAS macro, S-PLUS library and R package to perform logistic regression without convergence problems*. Retrieved from < http://www.meduniwien.ac.at/user/georg.heinze/techreps/tr2_2004.pdf > .

Henderson, J. M., Larson, C. L., & Zhu, D. C. (2008). Full scenes produce more activation than close-up scenes and scene-diagnostic objects in parahippocampal and retrosplenial cortex: An fMRI study. *Brain and Cognition, 66*(1), 40–49. http://dx.doi.org/10.1016/j.bandc.2007.05.001.

Henson, R. N., & Gagnepain, P. (2010). Predictive, interactive multiple memory systems. *Hippocampus, 20*(11), 1315–1326. http://dx.doi.org/10.1002/hipo.20857.

Hindy, N. C., Solomon, S. H., Altmann, G. T. M., & Thompson-Schill, S. L. (2015). A cortical network for the encoding of object change. *Cerebral Cortex, 25*(4), 884–894. http://dx.doi.org/10.1093/cercor/bht275.

Huff, M., Meitz, T. G. K., & Papenmeier, F. (2014). Changes in situation models modulate processes of event perception in audiovisual narratives. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(5), 1377–1388. http://dx.doi.org/10.1037/a0036780.

Huff, M., Schwan, S., & Garsoffky, B. (2011). When movement patterns turn into events: Implications for the recognition of spatial configurations from different viewpoints. *Journal of Cognitive Psychology, 23*(4), 476–484. http://dx.doi.org/10.1080/20445911.2011.541152.

Hymel, A., Levin, D. T., & Baker, L. J. (2015). Default processing of event sequences. *Journal of Experimental Psychology: Human Perception and Performance*. http://dx.doi.org/10.1037/xhp0000082.

Jackson, P. L., Meltzoff, A. N., & Decety, J. (2006). Neural circuits involved in imitation and perspective-taking. *NeuroImage, 31*(1), 429–439. http://dx.doi.org/10.1016/j.neuroimage.2005.11.026.

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences, 12*(2), 72–79.

Kurby, C. A., & Zacks, J. M. (2011). Age differences in the perception of hierarchical structure in events. *Memory & Cognition, 39*(1), 75–91. http://dx.doi.org/10.3758/

s13421-010-0027-2.
Lamm, C., Batson, C. D., & Decety, J. (2007). The neural substrate of human empathy: Effects of perspective-taking and cognitive appraisal. *Journal of Cognitive Neuroscience, 19*(1), 42–58.
Lea, C., Reiter, A., Vidal, R., & Hager, G. D. (2016). *Segmental spatiotemporal CNNs for fine-grained action segmentation*. ArXiv:1602.02995 [Cs]. Retrieved from < http://arxiv.org/abs/1602.02995 > .
Libby, LK., & Eibach, Rp. (2011). Visual perspective in mental imagery A representational tool that functions in judgment emotion, and self-insight. *Advances in Experimental Social Psychology, 44*, 185–245. http://dx.doi.org/10.1016/B978-0-12-385522-0.00004-4.
Libby, L. K., Shaeffer, E. M., & Eibach, R. P. (2009). Seeing meaning in action: A bidirectional link between visual perspective and action identification level. *Journal of Experimental Psychology: General, 138*(4), 503–516. http://dx.doi.org/10.1037/a0016795.
Loschky, L. C., Larson, A. M., Magliano, J. P., & Smith, T. J. (2015). What would jaws do? The tyranny of film and the relationship between gaze and higher-level narrative film comprehension. *PLoS ONE, 10*(11), e0142474. http://dx.doi.org/10.1371/journal.pone.0142474.
Loucks, J., & Baldwin, D. (2009). Sources of information for discriminating dynamic human actions. *Cognition, 111*(1), 84–97. http://dx.doi.org/10.1016/j.cognition.2008.12.010.
Magliano, J. P., Kopp, K., McNerney, M. W., Radvansky, G. A., & Zacks, J. M. (2012). Aging and perceived event structure as a function of modality. *Aging, Neuropsychology, and Cognition, 19*(102), 264–282. http://dx.doi.org/10.1080/13825585.2011.633159.
Magliano, J. P., Miller, J., & Zwaan, R. A. (2001). Indexing space and time in film understanding. *Applied Cognitive Psychology, 15*, 533–545.
Magliano, J. P., Radvansky, G. A., Forsythe, J. C., & Copeland, D. E. (2014). Event segmentation during first-person continuous events. *Journal of Cognitive Psychology, 26*(6), 649–661. http://dx.doi.org/10.1080/20445911.2014.930042.
Magliano, J. P., & Zacks, J. M. (2011). The impact of continuity editing in narrative film on event segmentation. *Cognitive Science, 35*(8), 1489–1517. http://dx.doi.org/10.1111/j.1551-6709.2011.01202.x.
Massad, C. M., Hubbard, M., & Newtson, D. (1979). Selective perception of events. *Journal of Experimental Social Psychology, 15*(6), 513–532. http://dx.doi.org/10.1016/0022-1031(79)90049-0.
Mcisaac, H. K., & Eich, E. (2002). Vantage point in episodic memory. *Psychonomic Bulletin & Review, 9*(1), 146–150. http://dx.doi.org/10.3758/BF03196271.
Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology, 28*(1), 28–38.
Newtson, D., & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology, 12*, 436–450.
Newtson, D., Engquist, G., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology, 35*(12), 847–862.
Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.
Nigro, G., & Neisser, U. (1983). Point of view in personal memories. *Cognitive Psychology, 15*(4), 467–482. http://dx.doi.org/10.1016/0010-0285(83)90016-6.
Peng, X., Wang, L., Wang, X., & Qiao, Y. (2016). Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding, 150*, 109–125. http://dx.doi.org/10.1016/j.cviu.2016.03.013.
Radvansky, G. A., & Copeland, D. E. (2006). Walking through doorways causes forgetting: Situation models and experienced space. *Memory & Cognition, 24*(5), 1150–1156.
Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science, 31*, 613–643.
Roche, K., & Chainay, H. (2013). Visually guided grasping of common objects: Effects of priming. *Visual Cognition, 21*(8), 1010–1032. http://dx.doi.org/10.1080/13506285.2013.851136.
Ruby, P., & Decety, J. (2001). Effect of subjective perspective taking during simulation of action: A PET investigation of agency. *Nature Neuroscience, 4*(5), 546–550. http://dx.doi.org/10.1038/87510.
Rust, N. C., & Stocker, A. A. (2010). Ambiguity and invariance: Two fundamental challenges for visual processing. *Current Opinion in Neurobiology, 20*(3), 382–388. http://dx.doi.org/10.1016/j.conb.2010.04.013.
Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*. http://dx.doi.org/10.1038/nn.3331 Advance online publication.
Schwan, S., & Garsoffky, B. (2004). The cognitive representation of filmic event summaries. *Applied Cognitive Psychology, 18*, 37–55.
Singh, S., Arora, C., Jawahar, C.V., (2016). First person action recognition using deep learned descriptors. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2620–2628, doi: 10.1109.CVPR.2016.287.
Smith, D. M., & Mizumori, S. J. Y. (2006). Hippocampal place cells, context, and episodic memory. *Hippocampus, 16*(9), 716–729. http://dx.doi.org/10.1002/hipo.20208.
Speer, N. K., Reynolds, J. R., Swallow, K. M., & Zacks, J. M. (2009). Reading stories activates neural representations of visual and motor experiences. *Psychological Science, 20*(8), 989–999. http://dx.doi.org/10.1111/j.1467-9280.2009.02397.x.
Speer, N. K., Swallow, K. M., & Zacks, J. M. (2003). Activation of human motion processing areas during event perception. *Cognitive, Affective, and Behavioral Neuroscience, 3*(4), 335–345.
Speer, N. K., Zacks, J. M., & Reynolds, J. R. (2007). Human brain activity time-locked to narrative event boundaries. *Psychological Science, 18*(449–455).
Sridharan, D., Levitin, D. J., Chafe, C. H., Berger, J., & Menon, V. (2007). Neural dynamics of event segmentation in music: Converging evidence for dissociable ventral and dorsal networks. *Neuron, 55*(3), 521–532. http://dx.doi.org/10.1016/j.neuron.2007.07.003.
Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science, 12*(2), 153–156.
Storms, M. D. (1973). Videotape and the attribution process: Reversing actors' and observers' points of view. *Journal of Personality and Social Psychology, 27*(2), 165–175. http://dx.doi.org/10.1037/h0034782.
Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences, 13*(9), 403–409. http://dx.doi.org/10.1016/j.tics.2009.06.003.
Swallow, K. M., Barch, D. M., Head, D., Maley, C. J., Holder, D., & Zacks, J. M. (2011). Changes in events alter how people remember recent information. *Journal of Cognitive Neuroscience, 23*(5), 1052–1064. http://dx.doi.org/10.1162/jocn.2010.21524.
Swallow, K. M., Zacks, J. M., & Abrams, R. A. (2009). Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology: General, 138*(2), 236–257.
Taylor, S. E., & Fiske, S. T. (1975). Point of view and perceptions of causality. *Journal of Personality and Social Psychology, 32*(3), 439–445. http://dx.doi.org/10.1037/h0077095.
Tversky, B., & Hard, B. M. (2009). Embodied and disembodied cognition: Spatial perspective-taking. *Cognition, 110*(1), 124–129. http://dx.doi.org/10.1016/j.cognition.2008.10.008.
Vallacher, R. R., & Wegner, D. M. (1987). What do people think they're doing? Action identification and human behavior. *Psychological Review, 94*(1), 3–15. http://dx.doi.org/10.1037/0033-295X.94.1.3.
Vannuscorps, G., & Caramazza, A. (2016). Typical action perception and interpretation without motor simulation. *Proceedings of the National Academy of Sciences, 113*(1), 86–91. http://dx.doi.org/10.1073/pnas.1516978112.
Vogeley, K., & Fink, G. R. (2003). Neural correlates of the first-person-perspective. *Trends in Cognitive Sciences, 7*(1), 38–42. http://dx.doi.org/10.1016/S1364-6613(02)00003-7.
Vogt, S., Taylor, P., & Hopkins, B. (2003). Visuomotor priming by pictures of hand postures: Perspective matters. *Neuropsychologia, 41*(8), 941–951. http://dx.doi.org/10.1016/S0028-3932(02)00319-6.
Wilder, D. A. (1978). Effect of predictability on units of perception and attribution. *Personality and Social Psychology Bulletin, 4*(2), 281–284. http://dx.doi.org/10.1177/014616727800400222.
Wilson, M., & Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin, 131*(3), 460–473.
Yordanova, K., Whitehouse, S., Paiement, A., Mirmehdi, M., Kirste, T., & Craddock, I. (2017). What's cooking and why? Behaviour recognition during unscripted cooking tasks for health monitoring. In *2017 IEEE international conference on pervasive computing and communications workshops (PerCom Workshops)* (pp. 18–21). http://dx.doi.org/10.1109/PERCOMW.2017.7917511.
Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science, 28*(6), 979–1008. http://dx.doi.org/10.1207/s15516709cog2806_5.
Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition, 112*(2), 201–216. http://dx.doi.org/10.1016/j.cognition.2009.03.007.
Zacks, J. M., Speer, N. K., & Reynolds, J. R. (2009). Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General, 138*(2), 307–327. http://dx.doi.org/10.1037/a0015305.
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*(2), 273–293.
Zacks, J. M., Speer, N. K., Swallow, K. M., & Maley, C. J. (2010). The brain's cutting-room floor: Segmentation of narrative cinema. *Frontiers in Human Neuroscience, 4*. http://dx.doi.org/10.3389/fnhum.2010.00168.
Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General, 130*(1), 29–58.