Perceiving Event Dynamics and Parsing Hollywood Films
James E. Cutting, Kaitlin L. Brunick, and Ayse Candan
Online First Publication, March 26, 2012. doi: 10.1037/a0027737

CITATION
A motion picture is composed of virtual events joined together.
—Gibson (1979, p. 297)

With any cut the viewer must make a very fast early "decision" as to whether it opens a different scene or event.
—Hochberg and Brooks (1996, p. 261)

Film—like language, music, and indeed life—can be parsed meaningfully into a hierarchy that has units within units within units (Shaw & Cutting, 1980; Tversky, Zacks, & Hard, 2008; Warren & Shaw, 1985). A film is normally considered to have three or four acts each lasting a bit less than 30 min (Thompson, 1999); about 40 to 50 scenes each lasting on average about 2.5 min (Bordwell & Thompson, 2004, p. 67; McKee, 1999); and in contemporary films, between 1,000 and 2,500 shots with mean duration of about 2 to 6 s (Cutting, DeLong, & Brunick, 2011; Salt, 2006). Within this hierarchy, some units have the psychological stature of being events. That is, viewers judge them to have beginnings, middles, and ends, with boundaries that are often denoted by changes in time and place, and that form separable segments within the ongoing audiovisual stream.

This article has two goals concerning film events and their relation to the broader psychological literature on events. First, we propose that an appropriate level of event analysis in film is the subscene. To foreshadow why we think the subscene is important, our results show that event units parsed in 24 films by eight viewers averaged about 115 per film, and about 1 min in duration. These values fall clearly between those of scenes and shots. Second, and more important, we demonstrate that parsing films into local event structure can be done reasonably well on the basis of visual information alone, without necessarily monitoring the intentions and goals of the actors. To be sure, we have no doubts that viewers always monitor intentions and goals, and can do so for multiple characters (Magliano, Taylor, & Kim, 2005). But to foreshadow why we think our results are important, we demonstrate that eight physical variables and a code for place/time change can account for 50% of the variance in the event parsing of Hollywood film. We will argue that the omnipresence of the physicality of change across subscene boundaries makes understanding popular film easier.

Our view contrasts obliquely with the wider psychological literature. More specifically, one finding concerning naturalistic events in unedited sequences is that, although at a scale of about 15 s in duration events are marked by greater physical change, for those events at a scale around 1 min, the function of perceived intention plays a greater role (Hard, Recchia, & Tversky, 2011; Newton, Engquist, & Bois, 1977; Zacks, 2004; Zacks, Speer, & Reynolds, 2009; Zacks, Kumar, Abrams, & Mehta, 2009). Indeed, Zacks, Speer et al. (2009, p. 308) suggested "the event perception literature has focused specifically on actors' goals and intentions as cues to event structure."

To accommodate our contrasting view, we propose that there is an important difference between perceiving naturally occurring events and perceiving edited film events. Having viewers parse...
videos of naturalistic scenes is to have them make sense of the unfolding part of life that they are observing; having viewers parse a film is to have them follow a carefully crafted story. We would claim further that there is no real story to a naturalistic scene, at least not one that noticeably contrasts with its description as a single stream of activity. The edited film event, and its role within the movie, is more complex and typically isolates one event within the multiple streams of narrative. In other words, an event in naturalistic video is a slice of life; an event in popular film is part of a highly stylized cultural artifact.

With respect to our first goal, the psychological literature on film events has focused on the scene (e.g., Magliano, Miller, & Zwaan, 2001; Zacks & Magliano, 2011; Zacks, Speer et al., 2009), not on what we call the subscene. Since scenes do not always have subscenes, and since subscene properties may be mimicked at a larger scale, we recognize that some confusion may arise. Nonetheless, we believe that the distinction between scenes and subscenes is psychologically important; the study of the physical properties of the subscene reveals the ways in which filmmakers control the attention of viewers and allow them to organize and remember events through manipulation of shot transitions, shot duration, shot scale, motion, luminance, and color. Moreover, many of these physical differences serve to signal changes in time and place, the oft-cited hallmark of scene changes. But before considering these visual variables further, we need outline the general theoretical framework of continuity editing as applied to events, offer some ideas from film theory, and then distinguish between scenes and subscenes.

Continuity, Discontinuity, and Events

Consecutive shots in live-action film are, by definition, taken from different camera positions. This fact has the potential to disorient the viewer, disrupting continuity. After a cut, a naïve viewer might ask herself: “What just happened? What connects what I just saw with what I am now looking at?” Indeed, adults who have never seen a movie before often give responses to film sequences that seem to reflect just such questions (Schwan & Ildirir, 2010; see also Balázs, 1970, p. 34).

During the silent film era, a number of filmmaking techniques were developed to overcome the potential visual disjointedness of shot-to-shot juxtapositions (Bordwell & Thompson, 2004; Smith, in press). These have been accepted by generations of film viewers, probably because they capitalize on how perception has always worked (Cutting, 2005): We didn’t evolve to watch films; instead, popular films have evolved to exploit our perceptual and cognitive capabilities and their foibles. To understand continuity is, in part, to understand those foibles (Smith, Levin, & Cutting, in press).

Continuity editing, the dominant form of an interlocking set of classical editing techniques (see, e.g., Berliner & Cohen, 2011), is created through handling space and time, giving shape to the narrative, binding shots within units, and separating one unit from another. Thus, the structure of most popular films shows both continuity and discontinuity in time and in space. What happens across shots within a unit often occurs within more or less continuous time and contiguous space; what happens across units often occurs across initially unknown gaps of time and space. As noted by Zacks and Magliano (2011, p. 442), “In a well-edited movie, cuts that classical film theory would identify as scene breaks should be identified as event boundaries, whereas continuity edits should not.”

Following Newton et al. (1977) and across many studies, Zacks and his colleagues have had viewers parse streams of film and video into events. For example, Magliano and Zacks (2011) modeled film events based on a widely accepted account of continuity (e.g., Bordwell & Thompson, 2004). They coded cuts in *The Red Balloon* (Lamorisse, 1956) as continuity edits (shots across cuts that were continuous in space, time, and action), spatiotemporal discontinuity edits (those discontinuous in space and/or time but continuous in action), and action discontinuity edits (those discontinuous in all three). They found that discontinuity of action was the strongest predictor of event boundaries, and concluded “commercial film editing is shaped to support the comprehension of meaningful events that bridge breaks in low-level visual continuity, and even breaks in continuity of spatial and temporal location” (Magliano & Zacks, 2011, p. 1). We concur.

### The Fabula and the Syuzhet

In fashioning their stories, filmmakers must overcome many constraints. Consider two. First, they almost always need to condense time—or more accurately, sample it over the chronology of the whole story behind the film. Alfred Hitchcock summarized one necessity for temporal trimming: "What is drama, after all, but life with the dull bits cut out" (Truffaut, 1983, p. 103). In Russian formalist film theory, the overall story of a film is called the fabula (e.g., Bordwell, 1985).

Second, having sampled from these larger, multichanneled streams of virtual life, filmmakers must stitch the samples together to create visual and sound material that makes sense, creates dramatic impact, and presents a smooth narrative flow. Again from Russian film theory, this single stream of shots is called the syuzhet. And, perhaps most importantly in our context, from many theoretical perspectives the viewer is thought to cognitively create the events in the fabula (she confabulates them) from the syuzhet as it is presented on the screen. For example, “In constructing a fabula, the perceiver defines some phenomena as events” (Bordwell, 1985, p. 51). From this perspective, it is the viewer who, in constructing the fabula, interprets the perceived goals and intentions of the characters (Bordwell, 2006; Zacks, 2004; Zacks, Speer et al., 2009; Zacks, Speer, Swallow, & Maley, 2010; Zacks & Swallow, 2007), and this top-down process is thought to help event segmentation. This view, we believe, assumes “the poverty of the stimulus.” This is a longstanding idea that exalts the importance of cognition over stimulus information. It had great currency in 20th

---

1 The second type of edit in Magliano and Zacks (2011)—a change in space or time without an action change—is common in *The Red Balloon* as the little boy is followed by the balloon around the Ménilmontant quarter of Paris. Our experience, however, suggests that it is not particularly common in most popular films.

2 Bordwell (1985) has endorsed the *syuzhet-fabula* distinction, but not general Russian formalist film theory. Instead, Bordwell (1989) has endorsed a cognitive film theory that is quite comfortable within contemporary cognitive psychology. The contrast here is that we are, perhaps more than Bordwell, aligned with an ecological approach within cognitive film theory (Anderson, 1996).
century accounts of language learning, but more recently has been known to have problems (e.g., Jackendoff, 2002). As applied to film, we think this idea underestimates the psychological impact of the physicality of film structure.

To be sure, when viewers parse an unedited naturalistic video or a simple animation (e.g., Heider & Simmel, 1944) they are free to impose structure on what they see. Indeed, part of the human condition is that people always stand ready to construct meaning from patterns, even meaningless ones. But in retort we assert that popular film is designed by filmmakers; it is not naturally occurring flow. Filmmakers have a deep understanding of the structure of the fabula and they craft the syuzhet with painstaking care. Given this, it would be surprising if the marks of their design did not completely penetrate all aspects of the presentation of the film, yielding signposts for event parsing.

In summary, our view is that the ongoing stream of a film can be parsed into events largely on the basis of physical attributes of the syuzhet; the understanding of actors’ intentions and goals may not be wholly necessary to this process. However, erecting the relations among those events, making those events mean— which is the first stage of constructing the fabula—is completely dependent on understanding those goals and intents. This new perspective, we suggest, forces a more thorough understanding of the physical variables that are part of the syuzhet, and should help us understand how and why the viewer parses events where they do.

**Scenes and Subscenes**

The notion of a scene comes from the theater. There and traditionally, a section of a play is considered a scene if it takes place in one location over a continuous but relatively small amount of time. Typically, if one character leaves the stage or another enters and if conversation or action continues, the scene is regarded as unchanged (Obstfeld, 2000; Polking, 1990, p. 405). This definition—the scene as an event occurring in a single location over continuous time irrespective of the population of actors— was used by Messaris (1994) to segment TV shows. Such a framework operates reasonably well in that medium, particularly for the sitcoms and soap operas that he investigated, and it can also be applied reasonably well to a number of films (e.g., Magliano et al., 2001).

Subscenes, which Bellour (1976) has called subsegments, mark somewhat finer boundaries by allowing for the entrances and exits of characters and changes in tone within a scene— neither of which are instances of change in time or place. But more importantly, subscenes also allow for the crosscuttering of events within parallel action, which demonstrate clear changes in time or place. Crosscutting occurs when a film follows, in parallel, two or more storylines with those narratives temporally interleaved. This technique has been common in films since D. W. Griffith in the 1910s (Bordwell, Staiger, & Thompson, 1985), if not before (Salt, 2009), and particularly in action and adventure movies. For example, in Rocky IV (Chartoff, Winkler, & Stallone, 1985) there is a 4-min sequence with Rocky Balboa (Sylvester Stallone) and Ivan Drago (Dolph Lundgren) doing strenuous physical training in vastly different locales. The action cuts back and forth between the two fighters 27 times over 174 shots, often with clear time lapses and location changes between depictions of the same fighter. Although viewers may differ, one might code this as one sequence (the preparation for the fight), two scenes (one with Rocky and one with Drago), and 28 subscenes each focusing on a single fighter and his entourage, each of which could be classified as an event.

**Corpus Analysis**

**Film Choice, Delivery, Viewers, and Variables**

A number of previous studies on the segmentation of dynamic events have used a single popular film. This is fine for promoting comparisons between naturalistic sequences and edited film, but with $n = 1$ and $df = n - 1$, this procedure cannot tell us much about the general event structure of popular film. For this reason, we assembled a relatively large corpus of feature-length films for a few observers to segment into events.

Previously, we parsed 150 popular films into their shots—10 films each from 15 different release years, every 5 years from 1935 to 2005 (Cutting, DeLong, & Nothelfer, 2010). To these we added 10 more films from 2010 (Cutting, Brunick, DeLong, Iricinschi, & Candan, 2011). We chose members of this 160-film corpus to exemplify five genres as determined generally by their listings on the Internet Movie Database (IMDb, http://www.imdb.com). These films were also among the highest grossing films of their release years, or before 1978, among those seen by the largest number of people reporting on the IMDb. From these 160 films we selected 24 films for further analysis—one drama, one comedy, and one action film each from release years 1940, 1950, 1960, 1970, 1980, 1990, 2000, and 2010. These are listed in the filmography (see Appendix). For our analyses, all films were downsampled to a frame size of 256 × 256 pixels.

Eight viewers, the authors and five undergraduate research assistants, worked on the films as part of a collective project. Three different viewers parsed each film. Viewers were instructed to segment each film into events. They were told that some accounts of events consider actions taking place at a single location within a single span of time, but they were specifically encouraged to determine on their own what an event might be in a given film. We did not stress the term scene, nor use the term subscene, because we felt the latter was too arcane and deviated too much from what we hoped would be the intuitive nature of the task. Nonetheless, it was stressed that an event could be no smaller than a single shot, that it would most often encompass a number of shots, that only rarely would it contain as many as 50 or more, and that it must begin and end with a shot transition—a cut, dissolve, fade, or wipe. This latter constraint allowed more control over measuring the physical variables than if we used a time-bin analysis, as done by Zacks and colleagues (e.g., Zacks, Speer, et al., 2009), and it also seemed more consistent with the integrity of edited film.

Viewers worked individually with frequent breaks. To familiarize themselves with the narrative, each viewer first watched each film with full sound and color (for the 18 films not in black and white) through to its end. They then parsed it on a second viewing. The whole task took between 4 and 10 hours per film. Films were played on laptops. Image size was about $8\,^\circ \times 8\,^\circ$ of visual angle. While parsing films, a viewer was able to stop the film and scroll frame-by-frame backward and forward to determine the frame number where the new scene began. In previous research (Cutting et al., 2010) and with subsequent rechecks, we recorded the frame
number of the beginning of each shot for each film on a spreadsheet. That file was open during the second viewing of the film so that the viewer could refer back and forth between the two. Each viewer marked the frame number of the first shot that began each new event. Viewers also often made notes concerning the content, which served as aids in remembering the flow of the narrative after taking a break. Perhaps most importantly, we decided in advance to analyze each individual’s parsing data separately. That is, we sought no parsing uniformity across viewers. We then averaged the results across viewers within a film and then across films.

**Interobserver Agreement**

Parsing concordance across the eight viewers and the three viewers of each film was reasonable, and probably as good as one could expect without interobserver consultation. For each shot, we coded whether one viewer indicated that it began a new event (1) or not (0), and whether another viewer indicated a new event or not (1 or 0). On 86.8% of all shots across all 24 films, the three pairs of viewers of each film agreed that no new event had occurred (0–0). On 5.7% of all shots they agreed that a new event had begun (1–1). And on 7.6% of all shots there was disagreement, with one viewer suggesting a new event had started and another not (0–1 and 1–0). Median interobserver agreement, $\kappa = .56$, was moderate across the 72 viewer pairs, and the mean agreement for each of the eight viewers against the others was $\kappa = .70, .61, .58, .56, .53, .53, .52,$ and $.45$. Two of these $\kappa$s are in the range of good agreement and six in the range of moderate agreement (Landis & Koch, 1977). We believe that these indices are not higher, in part, because Hollywood film is constructed on the backbone of continuity. Events generally glide together and one easily gets lost in the ongoing story of the film, being distracted from the task at hand (Smith & Henderson, 2008).

Most viewer disagreements occurred when one viewer gave a locally finer subdivision than another. More interestingly, about 10% of these occurred when one viewer thought that a given shot was the first of a new event whereas another thought it was the last of the previous. This often occurred when viewers felt that an event boundary occurred within a shot, had to decide which event to place it in, and decided differently. An example occurs about 56 minutes into *Harvey* (Beck & Koster, 1950) in the home Elwood P. Dowd (James Stewart). In the 50-s shot, an asylum attendant, Wilson (Jesse White), forms a personal relationship with Myrtle (Victoria Horne). As he is leaving the house the asylum doctor, Dr Chumley (Cecil Kellaway), enters without a cut and changes the tenor of the ongoing activities. Likely because this change occurs early in the shot, two viewers classified this shot as part of the event involving the doctor; but one did not, leaving it with the more romantic subscene.

There is also another kind of structural ambiguity. Consider a three-shot sequence about 20 min into *Erin Brockovich* (Hardy, Shamberg, & Soderbergh, 2000). The first shot shows Brockovich (Julia Roberts) at her desk looking over files, a dissolve occurs and then the second shot shows her hours later still at the desk looking over the files. There is then a cut to the third shot where Brockovich asks a coworker a question in another area. Dividing strictly by time, the second and third shots might belong to the same subscene, but dividing strictly by space the first and second shots might belong to the same subscene. Indeed, one viewer followed the first strategy, one followed the second, and a third thought all three shots were part of the same event.

**Subscene Analysis and Visual Variables**

The two theories of events most relevant to this research focus on changes. The event indexing theory (Zwaan, Langston, & Graesser, 1995) and the event segmentation theory (Zacks, Speer, Swallow, Braver, & Reynolds, 2007) both look toward unexpected or unpredicted changes in the flow of events to predict the presence of a new event. In that spirit, we have investigated the differences in physical attributes of shots that happen just before and just after an event boundary, and compare those to what happens before and after every other shot a the film that does not demarcate an event boundary.

Thus, after viewers parsed their films, we performed analyses on six physical attributes that are relevant to the visual dynamics of subscenes—for each film, we noted the locations of noncut transitions (dissolves, fades, wipes, etc.) with respect to viewer parses; we compiled profiles of the mean shot durations within viewer-parsed events, mean shot scales of all shots within those events, and mean motion and movement within those events; and we measured shot luminance and color differences within and between events. We hoped these visual variables would be correlated with changes in time and place, and in effort to assess that relation we also noted all changes in place or time in each film. We also analyzed several possible effects across release years and genres.

Of course, subscenes—and scenes and real-world events—come in different sizes. Here we measured their number of shots and their duration. Across the 24 films and three parsings of each film the mean, median, and modal number of shots per event were 11, 6, and 1 shot, respectively; and the mean, median, and modal duration of these events were 81, 57.5, and (grouped within non-overlapping 5-s bins) 15–20 s. Both distributions were essentially lognormal for each film. As a comparison, the mean durations for coarse- and fine-grain segmentations for a single film reported by Zacks, Speer et al. (2009, Experiment 2) were 51.5 and 13.8 s, respectively, and in Zacks et al. (2010) they were 59.4 and 22.9 s. Thus, it seems more likely that the subscene events in our films correspond more closely to the coarse-grain events in theirs.

**Scenes and shot transitions.** Consider first the traditional differences in how shots were separated across scenes. Carey (1974) noted that older popular films used dissolves, fades, and wipes for this purpose. Thus, the occurrence of one of these noncuts would seem to be strong information for the beginning of a new scene, if not a new subscene. Fades and wipes largely vanished from popular cinema after about 1970 but dissolves, although also less common, are still used (Cutting, Brunick, & DeLong, 2011). However, dissolves are also often strung together linking a number of shots in a Hollywood montage (Dmytryk, 1984; see also Salt, 2006), creating a particular mood over a condensation of time. Such devices make a dissolve less predictive of a scene change. Nonetheless, we tallied the occurrence and location of each noncut across each of film and sought assessment of how they aligned with event boundaries.

**Subscene shot-duration profiles.** Shot durations vary widely across and within films, but no systematic research has focused on the pattern of shot lengths within a film, or any smaller unit. It seemed obvious to us that patterns of shot lengths might play a role
in event segmentation in film viewing, and that they were worthy of investigation.

Previously, we had analyzed these films into their component shots and shot durations (Cutting et al., 2010; Cutting, DeLong et al. 2011). In our second analysis here, we wished to compare and average events across a film by resampling those that had more than one shot as determined by each parser. To understand our resampling method, consider first a representation of shot durations in three events in MASH (Preminger & Altman, 1970), shown in Figure 1. These were the 1st, 3rd, and 17th subscenes as parsed by one viewer. The first has six shots with a mean shot duration of about 28 s; the second has eight shots with a mean of about 16 s; and the third has 24 shots averaging about 6 s. To compare their shot-duration profiles we performed a simplified form of Procrustes analysis (Gower, 1975), which can be used to look for a common shape in datasets whose values differ in scale. Here, we affine transformed the one-dimensional array of shot durations. In particular, regardless of the number of shots in a film event, we resampled each subscene profile using linear interpolation, and assigned the resulting values to 200 equal-length consecutive bins. We chose this number because it was a bit larger than the largest number of shots in any event. Our procedure is suggested in the bottom panels of Figure 1. With each subscene resized to have the same number of bins, we could then compare all events within a film, and then all events across films.

**Subscene shot-scale profiles.** Next, we had been aware that many accounts of the relation between camera work and scene structure had focused on changes in shot scale (e.g., Bordwell & Thompson, 2004). That is, scenes often begin with a long shot showing the characters and their positions in space, and then hone in successive shots with medium close ups and close ups as the scene progresses. We set out to determine if this was generally true.

For this article, the authors coded the shot scale of every shot in each of the 24 films. We chose seven categories based on those of Bordwell and Thompson (2004, p. 262). That is, we categorized each shot as an extreme long shot (XLS = 1, where a standing human body as the focal object would be considerably smaller than the height of the frame), a long shot (LS = 2, where a standing person would just barely fit into the frame), a medium long shot (MLS = 3, where an individual is generally truncated at the knees), a medium shot (MS = 4, where the individual is cut off between the waist and chest), a medium close up (MCU = 5, where the height of the frame is mostly filled by the individuals head and neck), a close up (CU = 6, where the head typically extends slightly beyond the frame), to an extreme close up (XCU = 7, where only the eyes, a hand, or a small object might be shown). If no person were shown in the shot, these sizes were scaled to the focal object as it would compare to the human body.

Figure 2 shows a breakdown of these categories for a still frame out of the exploding café subscene in Inception (Nolan, 2010). The shot-scale data were then resampled in the same way as the shot-duration data shown in Figure 1, and then averaged across a parser’s subscenes, across parsers of a given film, and then across all films. Finally, we chose the initial scale of each shot because in many cases the camera dollies or zooms in or out, or pans to focus on people or objects near or far, making it too difficult to provide a single code that would be representative of the whole shot.

**Subscene motion profiles.** Zacks and colleagues have shown that motion is important for the segmentation of fine-grain events in naturalistic video, but less so in coarse-grain events. We already know that the time-scale of our events is more akin to Zacks’s coarse-grain events. Nonetheless, we felt it important to investigate the role of motion in segmenting Hollywood film.

We assessed the amount of motion of actors and objects and the movement of the camera as distributed across events using a method developed by Cutting, DeLong et al. (2011). There, we had correlated the luminance values of pixels in next-adjacent 256 × 256 pixel frames (e.g., frames 1 and 3, 2 and 4, 3 and 5, . . ., 32,707 and 32,709, etc.) along the length of each film. With correlation values stored for all pairs of frames for all films, we could then align them with the records of parsed events. As done previously, we subtracted these interface correlation values from 1.0 to create a visual activity index (1 − r). In this manner, values near zero (correlations near 1.0) signify very little motion and movement, and values increasingly above zero and ranging up to 2.0 signify increasingly more motion and movement. Changes across cuts were removed from the data. With visual activity values in hand, we then performed a procedure analogous to that for shot durations. That is, we interpolated those values along the course of each event, using the segmentation data of the viewer with the median number of parses for each film. As with the subscene shot-duration data, we created a uniform subscene length of 200 bins. We again then averaged those across subscenes within a film, and then across the 24 films.

**Shot luminance and color differences within and across subscenes.** Motion, of course, is but one of many possible visual variables that might contribute to parsing, so we decided to include two more—luminance and color. Luminance and color are not only logically and neurophysiologically independent, they also can be independent in the real world as well (Hansen & Gegenfurtner, 2009).

We first calculated the luminance for each shot in each film. If not already black and white, films were converted to grayscale such that all pixels in all frames of each shot had an 8-bit value, 0 (black) to 255 (white). We determined first the median of these values for each frame (Cutting, Brunick, DeLong, Iricinschi et al., 2011), and then the average across frames for each shot. For all

---

3 We recognize that these categorizations make ordinal data, and that to then treat them as if they have metric characteristics by rescaling and averaging them tempts a measurement theoretic violation. Nonetheless, these categories are based on continuous variation that occurs across shots and films, so we believe that the underlying data are continuous and that no harm is done by treating these data here as metric. Nonetheless, when logistic regressions are performed, we revert to treating these data as ordinal.

4 We chose not to use computer algorithms for measuring change across images. Standard models of change detection are focused on the recognition of objects that have undergone change (Radke, Andra, Al-Kofahi, & Roysam, 2005), not on measuring motion. We felt that our correlation technique was the simplest and reasonably represented the amount of motion and movement across frames. In addition, the reason for not choosing adjacent frames is that DVDs do not run at exactly 24 frames/s, whereas analog films do. If the conversion is not done carefully, which is the case for the DVDs of many of our older films, digital frames become hybrids of successive analog frames.
parsings, we then computed the difference in luminance for all adjacent shots within each event, and the luminance difference for all pairs of shots that crossed event boundaries. Finally, we compared the mean luminance differences for shots within and between events for all viewers and then all films.

Similarly, for the 18 color films in our sample, we measured the color differences across shots within and between events. To do this we converted each frame of each film, (an RGB jpeg file) into YCbCr format where Y is luminance, Cb is the blue–yellow opponent coordinate and Cr is the red–green opponent coordinate. We then averaged these values across all frames of each shot and then compared them. From the color coordinates in a two-dimensional space we measured the euclidean distance of that point from the two coordinates of the previous shot. This distance served as our measure of the overall color difference from one shot to the next.

Changes across subscenes in place, time, or both. Finally, because most accounts of film events begin with an emphasis on changes in place and/or time, we also coded these for each of the 24 films. Our overall argument is that most of the variables discussed above—shot transitions, shot durations, motion, luminance, and color—are low-level, physical attributes of films having nothing to do with content. Shot scale may need some recognition of the sizes of people, but beyond that recognition, it, too, is quite measurable with no knowledge needed for what is going on. The judgment of change in place or time, on the other hand, needs higher-level cognition. Nevertheless, they do not entail any judgment of actors’ intentions or goals.

Across-Corpus Results and Preliminary Discussion

Effects of Genres and Release Years

There were no reliable main effects and few interactions involving genre, although this could be due to the relatively small size of our corpus. In addition there were only a few reliable release-year effects. For example, the mean number of events in films across release years increased from about 80 in 1940 to about 150 in 2010 ($r = .54, p < .01$). Unsurprisingly, action films had an average of more subscenes (138.6) than comedies (100.8) or dramas (105.1);
We believe that the increase in the number of events per film argues for the increased complexity in their visual narratives.

In addition, since the length of films in our corpus did not change from 1940 to 2010 and since the number of events increased, one would expect that their mean duration would have systematically decreased across release years. And this is true ($r = -0.51, p < 0.02$), declining from about 90 s in 1940 to about 50 s in 2010. Although there were apparent differences in event duration among action films (49.6 s), comedies (78.1 s), and dramas (69.6 s), the variance was high across our corpus and there was no reliable pattern.

Finally, the mean number of shots per event increased across release years ($r = 0.66, p < 0.001$) from about 6 to about 15 from 1940 to 2010. This, too, is unsurprising since the median number of shots per film rose during this same period from about 500 to almost 2,500. Again, there were no reliable differences among action films (10.5 shots), comedies (10.0), and dramas (9.9).

Analytic Form for Further Analyses

We performed a logistics regression separately on the data of each film. The dependent variable was an ordinal code for the agreement among the three viewers about whether a new event had begun with the given shot. A code of 0 denoted complete agreement about a new event, a code of 2 indicated that two out of the three viewers agreed, a code of 1 showed that only one thought a new event had started, and code of 0 reflected agreement that all viewers thought that the current event was ongoing.

There were nine independent variables. The first eight we will call visual variables: (1) the transition type before the shot under consideration, with cuts given a nominal code of 1 and dissolves, fades, wipes, or other noncut transitions given a code of 2; (2) the duration (a continuous variable) of that shot; (3) the duration of the previous shot (also continuous); (4) the ordinal code for the shot scale of the current shot; (5) the ordinal code for the shot scale of the previous shot; (6) the change in the amount of motion in the current shot as compared to the previous shot (a continuous variable); (7) the change in the luminance and (8) the change in color of the current shot as compared to that of the previous (both also continuous). As suggested above, we also created (9) a nominal variable for change in place or time.

Our task was first to assess the relative potency of these physical variables to event segmentation, and then combine them with the place/time codes. Because we are leery of statistical fishing expeditions and because there is much more data here than in many empirical studies in psychology, we chose to analyze each film separately and use comparisons across films as cross-validation. Thus, for each film we created a matrix of variables for every shot (9 independent and 1 dependent), varying in length across the 24 films by the number shots, from just under 300 (Harvey; Beck & Koster, 1950) to nearly 3,000 (Mission: Impossible II; Cruise, Wagner, & Woo, 2000). Such differences have large effects on the magnitudes of the omnibus $\chi^2$ results of logistic regression, so below we report median $r_s$ as an effect size measure and the number of films for which the analysis was statistically reliable ($\alpha = 0.001$).

In brief, the visual variables discussed below, when taken in consort, accounted for a median of 29.8% of the variance in the parsing data across all films, and added to the place and time variable account for fully 50.2% of the variance. Table 1 also shows the median individual variance accounted for by each variable and the median rank for that variable as it accounted for variance in the regressions, one for each film. In addition, each logistics regression supports an overall measure of $a'$, the area under an ROC curve and a nonparametric measure of efficacy. The median effectiveness of the collection of eight visual variables in predicting universal parsing agreement among the observers (a dependent variable value of 3) was $a' = 0.897$ (maximum = 1.0).

When two observers agreed a new subscene had occurred the median was $a' = 0.850$, and when one observer parsed a scene at a given point in the shot stream the median was $a' = 0.801$. Since our data are nearly metric in quality for each film, each can be converted to a $d'$ for each film and medians taken. These are 1.85, 1.50, and 1.22, respectively. However, when the place/time code is added to the eight visual variables, the median $a'$ statistics soar to .976, .946, and .870, equivalent to $d'$s of 5.0, 2.70, and 1.65. Since $d'$ and Cohen's $d$ are identical, these are very large effects.

**Shot Transitions and Subscenes**

Dissolves, fades, wipes, and similar noncut transitions have not been common in popular films at least since 1930. Here, they make up only 5.4% of all transitions in this corpus for the 12 films released from 1940 to 1970. They are even less common in more contemporary films (here, 1980 to 2010) consisting of only 0.7% of all transitions. Because contemporary films have so many more shots than older films, however, the absolute number of noncuts is not quite so discrepant. Those films released on or before 1970 had an average of 43 noncuts (out of an average of 780 shots), and those released later had an average of 13 noncuts (out of 1,870 shots).

Logistics regression revealed that for 11 of 11 films in the earlier-release group, \textit{(MASH} [Preminger & Altman, 1970] has

<table>
<thead>
<tr>
<th>Information source</th>
<th>Event parsing variance accounted for</th>
<th>Median rank importance (out of 8) in regressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transition type</td>
<td>7.5</td>
<td>2.0</td>
</tr>
<tr>
<td>1970 and earlier</td>
<td>21.7</td>
<td>1.0</td>
</tr>
<tr>
<td>1980 and later</td>
<td>2.6</td>
<td>6.5</td>
</tr>
<tr>
<td>Shot duration</td>
<td>8.5</td>
<td></td>
</tr>
<tr>
<td>Current shot alone</td>
<td>5.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Previous shot alone</td>
<td>4.8</td>
<td>4.0</td>
</tr>
<tr>
<td>Shot scale</td>
<td>12.7</td>
<td></td>
</tr>
<tr>
<td>Current shot alone</td>
<td>10.2</td>
<td>2.0</td>
</tr>
<tr>
<td>Previous shot alone</td>
<td>2.7</td>
<td>5.5</td>
</tr>
<tr>
<td>Change in motion</td>
<td>1.1</td>
<td>7.0</td>
</tr>
<tr>
<td>Change in luminance and color together</td>
<td>7.2</td>
<td></td>
</tr>
<tr>
<td>Luminance alone</td>
<td>1.8</td>
<td>6.0</td>
</tr>
<tr>
<td>Color alone</td>
<td>4.4</td>
<td>4.0</td>
</tr>
<tr>
<td>Ensemble of eight visual sources</td>
<td>29.8</td>
<td></td>
</tr>
<tr>
<td>Place and time code</td>
<td>39.9</td>
<td></td>
</tr>
<tr>
<td>Place/time + eight visual sources</td>
<td>50.2</td>
<td></td>
</tr>
</tbody>
</table>
only cuts), there was a reliable effect of noncuts predicting event parses (median $r = .47$). For the later-release group, 8 of 12 films showed reliable effects but with much smaller effect sizes (median $r = .16$). For nine films, noncuts were the most potent source of parsing information but only one of these was released after 1970. There were two films from 1940 (Grapes of Wrath [Zanuck & Ford, 1940] and Santa Fe Trail [Wallis & Curtiz, 1940]), two from 1950 (Harvey [Beck & Koster, 1950] and The Flame and the Arrow [Hecht, Ross, & Tourneur, 1950]), three from 1960 (Inherit the Wind [Kramer, 1960], Ocean’s 11 [Milestone, 1960], and Spartacus [Douglas & Kubrick, 1960]), one from 1970 (Beneath the Planet of the Apes [Jacobs & Post, 1970]), and one from 1980 (The Empire Strikes Back [Lucas & Kershner, 1980]), the last because of its heavy and idiosyncratic use of wipes. Over all films, transition type accounted for a median of 7.5% of the variance in the parsed data. The median effect was 21.7% for the older films and 2.6% for the more contemporary ones, with a reliable decline over release years ($r = -.80, p < .01$).

### Shot-Duration Profile of Subscenes

We were interested first in the mean pattern of shot durations across film events. Previously, we had found that the shots just before and just after a dissolve, which in studio-era Hollywood films was used to separate scenes, were longer in duration than other shots (Cutting, Brunick, & DeLong, 2011). Thus, we were prepared to find something similar here. Figure 3 shows the averaged individual results for the parsings of six different viewers, two each for three films.

Notice first the scallop-shaped profile of all six parsings. Notice next the different ranges of shot durations for each of these films. Philadelphia Story (Mankiewicz & Cukor, 1940) has an average shot duration of 13.0 s, Beneath the Planet of the Apes (Jacobs & Post, 1970) one of 4.6 s, and Mission: Impossible II (Cruise, Wagner, & Woo, 2000) one of 2.5 s. This difference is completely consistent with reports of shot duration declining since about 1950 (Bordwell, 2006; Cutting, Brunick, DeLong, Iricinschi et al., 2011; Salt, 2006). More interesting, however, is that the same general scallop shape is roughly true of all films, regardless of average shot duration. And notice third that for the latter two films the number of events determined by one of the viewers was quite different than the other. Yet despite such differences, each plot shows a generally similar profile with longer duration shots at the beginning and end of each subscene, and shorter ones in the middle. Moreover, the shape correspondence across viewers is quite good. For the three films of Figure 3, the correlations of values across the 200 bins for these pairs of viewers are $r = .724, .625$, and .873.$^5$

The average of all 72 parsings (three viewers for 24 films) is shown in Figure 4, with mean shot duration plotted as a function of the proportion of time through the resampled subscene. The light gray areas indicate the 95% confidence intervals on the mean data. The mean correlation across the three viewers for each film and then across the 24 films was $r = .71$, and the median correlation across all comparisons of all viewers and all films ($n = 72*71/2 = 2,556$) was $r = .59$. Thus, the smooth scalloped pattern seen in Figure 4 is a reasonable representation of the shot-duration profile of a cinematic event. In particular, regardless of how an individual viewer parsed a film, on average the shots are generally longer at the beginning and end of a subscene and shorter in the middle. Zacks, Speer et al. (2009) suggested that film events have beginnings and ends; our data also show that, at least in terms of shot duration, these contrast with their middles.

The scallop in Figure 4 is apparent, but given the variable number of shots in subscenes, one should ask: What shots carry the weight of the initial and final tails of the plot? We can answer this is two ways. First and most straightforward, the median number of shots per parsed subscene was six. Thus, given the relative steepness of the beginning and end of the plotted data, it seems likely that only the first and last shots are systematically longer than the others. Second, we performed a simulation. Omitting parsed events with only one or two shots (<7% of all parsings), which generally would not affect the scallop pattern, and those with more than 30 shots (<6%), we coded events with 3 to 30 shots as having longer duration shots at the ends and uniformly shorter shots in the middle. Thus, a 3-shot event would be coded $[2,1,1,2]$, 4- and 7-shot events would be coded $[2,1,1,1,1,1,2]$, and so forth. These were then resampled to create values in 200 bins as before, and then each event-length token was weighted by the frequency with which it occurred across all viewers. For example, 3-shot events occurred in 9.1% of all parses, 10-shot events occurred in

---

$^5$ Given our resampling and combination procedure, the values in the 200 bins are clearly not independent. Moreover, it is not clear what the degrees of freedom should be for a statistical test. Thus, $r$ values are reported here simply as a general measure of congruence.

---

![Figure 3](image-url) The average resampled shot-duration profiles (in seconds) for subscenes in three films. All panels are the results of separate viewers. The number of subscenes parsed by each viewer is shown in the lower left corner of each plot. The curved lines represent the best quadratic fits to the data. See also Figure 1.
3.2%, and 30-shot events in 0.8%. The weighted average result is shown in the inset to Figure 4. Clearly, these three simulated data are entirely consistent with the idea that only the first and last shots of a film event are relatively long, with those in between generally shorter and reasonably uniform in duration.

In logistic regressions, the durations of the current and the previous shots were reliable predictors of a new event in 20 of the 24 films for each measure (median \( r_s = .22 \)). Only in Harvey (Beck & Koster, 1950) did both fail to reach significance. The current-shot duration accounted for a median of 5.0% of the variance in the parsed data. Indeed, it was the most potent source of information for five films: All About Eve (Zanuck & Mankiewicz, 1950), MASH (Preminger & Altman, 1970), Nine to Five (Gilbert & Higgins, 1980), Goodfellas (Winkler & Scorsese, 1990), and Home Alone (Hughes & Columbus, 1990). The last shot of the previous subscene accounted for 4.8% of the variance, and it was the most potent source in two films—Five Easy Pieces (Rafelson, 1970) and Mission: Impossible II (Cruise, Wagner, & Woo, 2000). Together and because they are correlated, as would be suggested in Figure 4, the shot durations of the first and last shots of subscenes accounted for somewhat less variance than their sum (9.8%), a median of 8.5%.

### Shot-Scale Profile of Subscenes

As with the series of shot durations in each event, the series of shot scales (the ordinal values shown in Figure 2) for each event was affine transformed and interpolated into 200 bins. The resulting profiles were then averaged across each observer, across the three observers of a given film, and then across films. Overall results are shown in Figure 5, where the line of solid thick dots represents the mean profile across the 72 parsings and the gray areas the 95% confidence intervals. Notice that, as suggested by Bordwell and Thompson (2004), the initial shot, or shots, of a subscene tend to include more background than subsequent shots. However, against the Bordwell and Thompson account there is a slight tendency for a final shot to do so as well.

We found that 21 of 24 films showed reliable effects of current shot scale predicting parses (all but Inherit the Wind [Kramer, 1960]; MASH [Preminger & Altman, 1970]; and Beneath the Planet of the Apes, [Jacobs & Post, 1970]), it accounted for a median of 10.2% of the variance in the data (median \( r = .32 \)), and was the most potent source of information in six, mostly more recent, films: Philadelphia Story (Mankiewicz & Cukor, 1940), Die Hard 2 (Gordon, Gordon, & Harlin, 1990), Erin Brockovich (Hardy, Shamberg, & Soderbergh, 2000), What Women Want (Meyers, 2000), Valentine’s Day (Karz, Rice, Rosen, & Marshall, 2010), and The Social Network (Spacey, 2010). Moreover, 15 of 24 films showed a reliable effect of the previous shot size also predicting a new event, accounting for 2.7% of the variance in the data (median \( r = .16 \)), although it was never the most potent source of information for a given film. Together, overall shot scale information (current shot and previous shot) accounted for 12.7% of the variance in the parsed data, about the same as their sum (12.9%).

The effect of the scale of the last shot in a subscene is interesting, and somewhat against the received view in film theory. For many film-event boundaries, the change in shot scale seems at least as important as the direction of change. For example, about 5 minutes into Goodfellas (Winkler & Scorsese, 1990), the film’s protagonist, the younger Henry Hill (Christopher Serrone), disappears in a long shot as he runs away from his parents’ house, ending one event. It is followed by a cut to a close-up of hands around whiskey bottles as gang members bribe policemen, and as the young Henry looks on. All three viewers chose this as an event boundary, despite the fact that the soundtrack is a continuous voice over. In this manner, a number of subscenes may end with a long shot as the next subscene begins with a close-up, the reverse of the more classic pattern.

Another release-year effect is found in shot scale. Bordwell (2006) suggested that more contemporary films use more close-ups...
than early films, and Salt (2006) found that this was particularly true for films from the early sound era up through the 1950s. Among our 24 films from 1940 to 2010, we also found increasingly tighter shots in more contemporary films \((r = .45, p < .02)\). That is, in our 1940s films, the average shot scale was close to a medium-long shot (3 in Figure 2), whereas for those from the 21st century, it is closer to a medium shot (4 in Figure 2).

**Motion Profile of Subscenes**

Next, we focused on the motion and movement across the average event. We chose the data of one parser from each of the 24 films, that of the viewer who marked the median number of events. Then, from the correlations done previously on the amount of frame-to-frame visual activity across the length of each film (Cutting, DeLong et al., 2011), we grouped the activity in those frames, fitting them to each parsed event. Again, we affine transformed the data to fit each subscene into 200 bins (ignoring shot boundaries). We then calculated the median visual activity value within each bin. Events with fewer than 200 frames (8.33 s) were ignored (<7%). We then averaged the resulting values across events within a film, and then across the 24 films. The result is shown in Figure 6, with visual activity plotted as a function of the proportion of time through the subscene.

Although much noisier, the overall pattern in Figure 6 is somewhat similar to that in Figure 4. That is, after a slight increase in motion over the first 5% of the event, there is a downwardly bowed scallop that follows. The data for each film were fit with a quadratic equation and 23 of the 24 films showed this downward bowing pattern (the exception was Five Easy Pieces, [Rafelson, 1970]). Each of the first 47 bins shows more activity than each of the 153 bins that follow. There is also a slight increase in motion at the tail end of the events; bins 160–200 have more motion and movement than bins 110–150 \((d = 2.05)\). Unsurprisingly, contemporary films have a lot more motion in them than do films from the studio era (Cutting, Brunick, DeLong, Iricinschi et al., 2011), but the same general pattern seen in Figure 6 was true for most films. Again, the 95% confidence limits on the means are shown in gray.

Because of the increased noise in the data and because this effect is relatively small, we assessed the effect of motion and movement in parsing simply by measuring the difference in median visual activity indices for consecutive shots. The change in such activity from one shot to the next reliably predicted scene parses for 10 of the 24 films, but it accounted for only a median of 1.1% of the variance in the data \((median \ r = .11)\). Moreover, it never ranked higher than the third most potent source of parsing information in any film (for Inception [Nolan, 2010]).

**Luminance and Color Changes Across Subscenes**

We considered next the variation in general brightness within and across film events. The mean within-event across-shot luminance difference for the 24 films was 12.6 (SEM = 0.63) on the 8-bit \((0–255)\) scale. The mean between-event across-shot difference was 20.9 (SEM = 1.47, \(d = 1.5\)). Although older films tend to be brighter than contemporary films (Cutting, Brunick, DeLong, Iricinschi et al., 2011), this overall relative difference was reliable for all but one film (Harvey, Beck & Koster, 1950).

Figure 7 shows examples of shots and event boundaries from the film Inception (Nolan, 2010) that follow this pattern. All three viewers agreed that the boundaries between shots 55 and 56 and between shots 66 and 67 parsed this stream of shots. Note that these show a generally larger luminance differences, dark to light or light to dark, than do the other pairs of shots within subscenes. Only for Inception was luminance change the best predictor of parsing among the seven variables. Overall, the ordinal logistics regression showed that luminance differences reliably predicted parses for 16 of the 24 films, but only accounted for a median of 1.8% of the variance in the data \((median \ r = .13)\).

Differences in color were more potent than luminance as a parsing predictor. For 17 of the 18 color films it reliably predicted a change in events. It accounted for 4.4% of the variance \((median \ r = .209)\), but was the most salient source of parsing information only for Ordinary People (Schwary & Redford, 1980). In that film, teenager Conrad Jarrett (Timothy Hutton) tries to suppress a near-nightfall sailing accident in which his older brother was killed. That event is relived in flashbacks that are tinted blue whereas the ongoing day-to-day events, particularly those in the office of the therapist (Judd Hirsch), are mostly brown.

For each film, we normalized to 1.0 the mean color space \((C_{bC'}r)\) distance for all within-event across-shot differences, and compared the magnitude of those differences to the between-event across-shot differences. This yielded a mean value of 2.76 and a highly reliable difference seen across all 72 comparisons \((d = 3.2\)).

Finally, luminance and color together accounted for 7.2% of the variance, a bit more than their sum \((6.2\%)\) and showing that they are negatively correlated in these films.

To understand better how differences in color (and luminance) can be used, again consider Inception (Nolan, 2010). A good part of the structure of the film takes place within four nested dream states. Level 1 takes place in a gray, rainy, urban setting; Level 2 within a hotel with brownish tones; Level 3 in a wintry environment around and inside a generally gray fortress/hospital; and
Level 4 (limbo) mostly inside brown urban interior. Figure 8 shows brightened versions of the average of every frame from the shots within those four dream states, along with sample stills from those levels. One can see that Levels 1 and 3 are largely gray, and Levels 2 and 4 are largely brown. These differences could aid viewers in recognizing where they are in the story that has over 400 jumps between levels.

Place and Time, and the Addition of the Visual Variables

As noted above, the basic definition of an event starts with place and time. Indeed, Messaris (1994) and others have used changes in one or both to discuss scene changes. Our original view was that the eight visual variables might serve as surrogates for place and time as markers of event boundaries. Thus, with the codes for time and/or place change (0 = no change; 1 = change) for each shot, we could predict the parsing strength (0, 1, 2, or 3) for each shot for each film. Impressively, this place/time code by itself accounted for a median of 39.9% of the variance in the data across all films. Such results are entirely consistent with the event models approach of Zacks et al. (2007), in which observers, as they parse events, track changes in time and in space.

To check on the correspondence of the eight visual variables with the place and time code, we entered all of them into new logistic regressions. The median variance accounted for was now 50.2%. In other words and somewhat to our surprise, about 10% of the variance remained as the contribution from the visual variables. Mostly this was from initial shot durations (still reliable in 13 films), final shot durations (13 films), and initial shot scale (9 films). In retrospect, however, these make sense. A number of subscene boundaries reflect the exits and entrances of characters, as Bellour (1976) suggested. Given that these take place in the same locations and in adjacent time, the place/time code would not change, and the motion, luminance, and color are not likely to change either. These results also suggest that something about event place and time change remains outside the purview of these visual variables. The 10% superiority of the place/time code over the current ensemble also suggests that some of this residual might be addressed by music or by sound effects, which are beyond the scope of this investigation. Nonetheless, we are quite impressed that 50% of the variance in the parsed data is accounted for by the collective information about place and time in conjunction with the physical information about transitions, shot durations, shot scale, motion change, luminance change, color change. We attribute the magnitude of these results to the intents of filmmakers in crafting their syuzhet.

Psychological Functions of Visual Information in Subscenes

The Establishing Shot

To contribute to the narrative of a play, a novel, or a film, a scene must perform one or more of three functions: develop a character, advance the plot, or give a causal analysis of why something is happening (Obstfeld, 2000). In popular film, a scene must first be established, and a subscene returning back to a previous scene must be reestablished. Following Hochberg and Brooks (1996) quoted in the epigram, the viewer must quickly come to know if a new event is occurring, where the action is taking place, and who’s doing it; or recognize that the new event has returned to pick up a previous one.

Setting the location of a new scene is often done with an exterior shot of a building within which the subsequent action will take place. This is called an autonomous shot (Bellour, 1976; Metz, 1974), and the camera often uses a telephoto lens. But more commonly a scene may “begin with a long shot that establishes the total space” (Bordwell et al., 1985, p. 63). In film practice, long shot refers to shot scale, not to shot duration. Such an establishing shot “must immediately reveal two things about the characters: their relative positions and their states of mind” (Bordwell et al., 1985, p. 63). The latter is revealed by facial and vocal expression, body posture, and is of course aided greatly by the content of what
has gone before. Our results confirm, as shown in Figure 5, that most subscenes begin with a longer-scaled shot than average.

In addition, although all shots in a film present new information, the establishing shot in a scene or reestablishing shot in a subscene must present more information than most. Quite simply, this demands time; the mental gears of the viewer need to be shifted to a new context. This mental shifting in film viewing has a parallel in reading. Readers spend an increased amount of time with text at the beginnings and ends of written scenes (Zwaan, Langston et al., 1995; Zwaan, Magliano, & Graesser, 1995). They read more slowly because they need more time and resources to digest the segments within the written narrative and to prepare themselves for new upcoming events. This may account, at least in part, for why (re)establishing and ending shots are generally about 50% longer in duration than most other shots in that event.

Motion is another source of information for subscene change. Indeed, Dmytryk’s Rule 5 of editing (Dmytryk, 1984, p. 38) might have predicted the general pattern seen in Figure 6: “All scenes should begin and end with continuing action,” implying that the middle of a subscene might have less motion. This idea is entirely consistent with those of Newtonson et al. (1977) and Zacks, Speer et al. (2009), who found that people tended to identify event boundaries when an actor’s position was changing rapidly, and with those of Zacks (2004), who reported that increases in the motion of simple animated objects were associated with event boundaries. Nonetheless, Zacks and colleagues have found motion to be most relevant for fine-grain events, and the film events that we’ve analyzed here are at the time scale of his coarse-grain events, where motion was less relevant. Our data show that at the beginnings and ends of subscenes the action is somewhat greater than what goes on during the middle, but the impact of this source of information is negligible on the parsing data. Its relative weakness across our corpus suggests that filmmakers long ago discovered other and more effective means for delivering information about event boundaries.

Finally, changes in luminance and color play a role in event parsing in film, as suggested in Figures 7 and 8. From our data, and from Hochberg and Brooks (1996, p. 261) who speculated that “lighting levels, sound track, and so forth are obvious factors” in event segmentation, the generally increased luminance and color differences across shots at subscene boundaries can serve that purpose. After all, different scenes generally occur in different locations in different types of lighting and often at different times of day.

The Remainder of a Subscene

After the scene has been established, or subscene reestablished, it must then develop in a causal way. “Characters act toward their goals, enter into conflict, make choices, set deadlines, make appointments, and plan future action” (Bordwell et al., 1985, p. 64). Notice the arc of the scene; tension is incrementally built and then released at the end (Pearlyman, 2009). This view seems at odds with more standard psychological event theory (Zwaan, Langston et al., 1995; Zacks et al., 2007), where the lack of predictability triggers an event boundary. The notion is that, instead of the end of an event being something of a surprise, the end of a film event is anticipated, tension rising until it is released.

Our results suggest that this tension is partly conveyed by the increased pacing of shots after the establishing shot (shot durations are shorter), and by adjustments in shot scale, turning more to close-ups so that emotions and facial expression can be better discerned. Motion and movement also diminish, and luminance and color are more uniform and generally consistent with that of other shots in the subscene. In this way, the narrative development tailors the presentation of the shots and dictates their physical form; and, in reverse, the physical variables of shots presage narrative development, helping to predict its continued path.

Narrative development within a scene increases tension. It weaves threads of causal action, some new and some old (from previous scenes), until a “causal line...motivates the shift to a new scene. The famous ‘linear’ of classical Hollywood cinema thus consists of a linkage which resembles a game of dominos, each dangling cause matched by its effect in the following scene” (Bordwell et al., 1985, p. 64; see also Thompson, 1999). The key idea here is the dangling cause (a clear pun on one kind of ill-formed sentence). It often occurs in the last shot of a scene. The dangling cause is the motivating link from one scene to the next, or to one later in the film. However, unlike the establishing shot, the final shot is most often a medium shot or even a close-up, long in duration but not reliably different in shot scale than most of the previous shots in the subscene. Nonetheless, it often “carries us to the heart of the space, the site of the character interaction, and leaves us there to anticipate how the dangling cause will be taken up later” (Bordwell et al., 1985, p. 66).

Our data suggest at least two ways in which this tension of anticipation is increased and moves toward wrapping up a subscene. First, and more generally, dangling causes either take extra time to motivate, or it may use extra time for emphasis. This extra shot duration can signal, by itself, an upcoming change. Second, shot scale can be used to provide anticipatory information. Tension mounts when the camera moves closer and closer to the characters, alternating in a shot/reverse-shot sequence (focusing first on one character then the other in a repeating sequence). Occasionally, as shown in the slight downturn of the end of subscene shot-scale data in Figure 5, the camera backs off. In such cases the dangling cause may already have been delivered, and tension somewhat released. This too can serve as information for anticipating change; the wrapping up has already begun.

We emphasize, however, that a film viewer need not be overtly aware of any of these cinemetric variables while watching a film. We suggest instead, and since Hollywood style is often called “invisible style” (Messaris, 1994), that filmmakers subtly manipulate all of these physical cues in the construction of their products to shape visual understanding (Smith, in press) and to increase emotional involvement.

Summary and Conclusion

Like text, music, and most of life, popular film can be organized into a hierarchy of units. Film typically has frames within shots, shots within subscenes, subscenes within scenes, scenes within sequences, sequences within acts, and acts within the totality of the film. Subscenes, the central unit of this investigation, have a set of physical signatures that we think strongly aids the viewer in parsing the overall narrative. In general, they have a scallop-like profile of shot durations, with longer duration shots at the begin-
ning and end and shorter shots in the middle; they begin with shots that include more of the surroundings than subsequent shots; they begin with generally larger differences in the amount of motion from the shot at the end of the previous scene; and they begin with a shot that contrasts in luminance and color by an amount larger than adjacent shots within a subscene. Shots at the end of sub-scenes are also long in duration, have slightly in increased motion, and deviate in shot scale from what will happen next.

Subscene onset and offset signatures appear to have several psychological functions. First, longer duration establishing shots allow the viewer more time to understand both the location of the actions and the mental states of the actors. Second, longer duration final shots allow the viewer more time to understand the dangling causes that will carry information forward for understanding subsequent scenes and subscenes. Third, the changes in shot size generally progress from the viewer being at some distance from the action generally toward being much closer to it, although sometimes the progression is the reverse. Fourth, the change in motion in film generally reflects how people parse the events in the world, although the effect in edited film seems not nearly as strong as in other contexts. Fifth and sixth, the changes in luminance and color are available to mark breaks in continuity, directing viewers to digest the previous subscene and get on with the understanding of the current one.

Finally, about 30% of the variance in the parsing data is accounted for by visual variables in the ongoing visual stream (the *syuzhet*) of the film, without considering physical aspects of the sound track, especially sound effects and music. In addition, this burgeons to about 50% when a simple code of a place or a time change is added. Notice, however, that we’ve said nothing about the actual content of the narrative, or the intentions and goals of the actors. It is our view that the constraints of storytelling in popular film are such that the physical variables in the *syuzhet* and the structure of the narrative in the *fabula* are deeply enmeshed, and that filmmakers have designed their films to be that way. In other words, parsing the structure of popular film is strongly aided by the physicality of the film. We suggest that this physicality is highly correlated with, and aids, the interpretation of events and goals of actors by film viewers, which has been the focus of nearly all previous research on film events.

References


**Appendix**

**Filmography and Genre Classifications**

**1940**
- *Santa Fe Trail*, action (Wallis & Curtiz, 1940).
- *The Grapes of Wrath*, drama (Zanuck & Ford, 1940).

**1950**
- *Harvey*, comedy (Beck & Koster, 1950).
- *The Flame and the Arrow*, action (Hecht, Ross, & Tourneur, 1950).
- *All About Eve*, drama (Zanuck & Mankiewicz, 1950).

**1960**
- *Inherit the Wind*, drama (Kramer, 1960).

**1970**
- *Beneath the Planet of the Apes*, action (Jacobs & Post, 1970).

**1980**
- *Nine to Five*, comedy (Gilbert & Higgins, 1980).
- *Ordinary People*, drama (Schwary & Redford, 1980).

**1990**

**2000**
- *Erin Brockovich*, drama (Hardy, Shamberg, & Soderbergh, 2000).

**2010**
- *The Social Network*, drama (Spacey, 2010).
- *Inception*, action (Nolan, 2010).

**Other Films**
- *Rocky IV*, action (Chartoff, Winkler, & Stallone, 1985).
- *Gigi*, comedy (Freed & Minelli, 1958).
- *The Red Balloon* (*Le Ballon Rouge*), fantasy (Lamorisse, 1956).

Received July 31, 2011
Revision received February 3, 2012
Accepted February 3, 2012