HUMAN–COMPUTER INTERACTION, 2003, Volume 18, pp. 325–372


# Using Film Cutting Techniques in Interface Design

**Jon May and Michael P. Dean**
*University of Sheffield*

**Philip J. Barnard**
*MRC–Cognition and Brain Sciences Unit*

## ABSTRACT

It has been suggested that computer interfaces could be made more usable if their designers utilized cinematography techniques, which have evolved to guide the viewer through a narrative despite frequent discontinuities in the presented scene (i.e., cuts between shots). Because of differences between the domains of film and interface design, it is not straightforward to understand how such techniques can be transferred. May and Barnard (1995) argued that a psychological model of watching film could support such a transference. This article presents an extended account of this model, which allows identification of the practice of collocation of objects of interest in the same screen position before and after a cut. To verify that filmmakers do, in fact, use such techniques successfully, eye movements were measured while participants watched the entirety of a commercially

**Jon May** is a psychologist with an interest in the application of unified models of cognition to perception and emotion, particularly with regard to the effects of task and context; he is a senior lecturer in the Department of Psychology at the University of Sheffield. **Michael Dean** is a psychologist who has researched the mental representations that give rise to perceptual and memory effects in object perception; he is now working as a speech therapist. **Philip Barnard** is a psychologist with an interest in theories of mental architecture and their application to complex tasks, emotion, and a range of psychopathologies; he is on the staff of the Medical Research Council's Cognition and Brain Sciences Unit.

CONTENTS



released motion picture, in its original theatrical format. For each of 10 classes of cut, predictions were made about the use of collocation. Peaks in eye movements between 160 and 280 msec after the cut were detected for 6 of the 10 classes, and results were broadly in line with collocation predictions, with two exceptions. It is concluded that filmmakers do successfully use collocation when cutting in and out from a detail, following the motion of an actor or object, and in showing the result of an action. The results are used to make concrete recommendations for interface designers from the theoretical analysis of film comprehension.

## 1. FILM AND COMPUTER INTERFACE DESIGN

This article argues that computer interface design can benefit from aspects of filmmakers' craft knowledge, but that identifying which aspects are beneficial, and how they can be applied, is not straightforward. To do so, we develop a general cognitive model of the perception of dynamically changing scenes, using examples from filmmaking, which emphasizes the need to present visual information in a manner that allows the all-important primary task (in film, narrative comprehension; in human–computer interaction [HCI], achieving task goals) to be processed without interruption by secondary tasks created by the need to repeatedly reorient to a changing visual scene. To test the assumption that this practice is actually successfully employed by filmmakers, we present eye-movement data collected from volunteers watching a commercial film. We illustrate the way that this par-

ticular principle is currently violated in interface design and how it could be usefully applied.

One of the major problems facing the designers of computer interfaces is that of "screen real estate," the limited amount of physical space available for the display of information about the computer user's task. The development of windowing systems in the 1980s promised a solution of sorts by defining tiled or overlapping rectangles of the screen so that information relevant to the current, most important aspects of the user's task could be presented in a foreground window with other task elements partially hidden or relegated to background status in other windows. The user can carry on the main task in the foreground window, interleaving other subsidiary tasks at will by activating background windows. The operating system or active software may create new windows to present task-relevant information or to request user input, such as the destination file name following a user's selection of the Save command.

In the perfectly designed HCI, all of these windows would open and close appropriately, displaying information exactly when users needed it, in such a manner that they could immediately comprehend both its content and its relevance to their ongoing task. Should users need to select a function, they would know exactly what the icon looked like and where to find it on the screen, or within which pop-up or pull-down menu it might be located, so that they would not have to stop doing their primary task to indulge in time consuming searches of the interface and of their memory.

A frequent complaint of computer users, in contrast, is that the interfaces they have to use are overcomplicated, or confusing, and that they neither know what many of the icons and screen objects mean or do, or why they come and go. When new windows open, even if the user is expecting them, the information within them may not be arranged appropriately, so the user has to actively search for the relation between the new and the previous view. In this regard, computer interfaces have been compared disparagingly with cinema films (e.g., Young & Clanton, 1993) in which film directors and editors frequently cut between shots to present new views of a scene, of different scenes, and even of action occurring at widely separated places and times. People rarely emerge from cinemas baffled by film cutting, complaining that they did not know where to look or that they missed crucial elements of the narrative because they were trying to work out what the new shot meant.

Of course, this comparison is unfair in several ways. The computer user is frequently engaged in several, more complex tasks than the viewer of a film, who usually has just one task to perform (the comprehension of a narrative that is usually contiguous and cumulative, guiding them through the film). The progress of the film is completely under the control of its design-

ers, and the viewer is passive in that they cannot change the film or alter its rate of progress, whereas the progress of the HCI is largely under the control of the user, and the designers of an application have little control over the context within which windows will need to be opened, or on the information that they may need to contain. However, it must be emphasized that an interface event is never directly caused by the user but is always a result of a design decision. The fact that the cursor arrow appears on screen in a particular place is not because the user has moved it there with the mouse, but because designers have programmed their application to interpret the hand movements detected by the mouse in a particular way. Every single interface event has been designed—and needs to be designed well—in the same way that every frame of a film has been constructed by the film's makers.

However, some of the differences between film and interface design should work in the computer user's favor, making interfaces easier to comprehend than film. The computer user is focusing on a specific train of information, whereas the film viewer may have to actively search the screen to deduce which information is relevant. If a computer window opens in response to some action that the user has performed, then the very fact that it is related to their task means that they will have certain predictable semantic or thematic knowledge that the designers can capitalize on. By presenting the information in the new window in a predictable physical position with relation to the relevant information in the old window, the user will be able to see and comprehend it without having to engage in an additional search task.

This principle, which May and Barnard (1995) called *collocation,* is employed by filmmakers when they construct match cuts. Simply put, the film is often cut so that the object that the filmmaker wants or expects the viewer to attend to in the new shot is placed close to the position of the object that the viewer can be expected to have been attending to in the previous shot. This is just one of many strategies filmmakers use in making their products comprehensible, but as it is based on physical location rather than narrative or thematic relations between objects, it is potentially the most directly applicable to interface design. May and Barnard (1995) presented an initial account of the cognitive tasks involved in watching film, where *filmic* cuts that did not interrupt the viewer's propositional and implicational processing of the narrative were distinguished from *unfilmic* cuts, which required the viewer to attend to the object structure of the scene to relocate items of thematic relevance. *Match cuts* are an example of filmic cuts; an example of unfilmic cuts are *jump cuts* in which an attended object's position or relation to another object is altered. May and Barnard (1995) argued that computer designers could improve the usability of their interfaces by employing filmic strategies, and

avoiding unfilmic strategies, to support narrative and perceptual coherence in the interaction.

That current interfaces do not generally follow such simple principles is easy to demonstrate. Using your favorite graphics application, open a large image and select the tool that allows you to zoom in on a detail (often represented by a magnifying glass or loupe icon). Click on a detail in any corner. If the interface uses the filmic principle of collocation, the detail should be displayed in the same corner as it was before you clicked it (albeit larger). Almost all applications actually move it to the center of the screen, leaving you looking at some other detail. You have to relocate the detail that you operated on. At least you can, with practice, learn that it will be centered. Zooming out is implemented in even less predictable ways across different applications (and many users do not know how to carry out the operation, which usually involves the same tool and a simultaneous key press).

The central argument of this article is that the design of computer interfaces could be improved if the interfaces made use of filmic principles, and in particular, collocation of the thematic topic of the user's task, whenever the scene portrayed on the interface changes. At a more abstract level, it is possible that the coherence of an interface could also be improved by adopting filmic devices that have been developed to convey a sense of narrative structure across shots and scenes, but that is beyond the scope of this article. Here we concentrate on the organization and dynamic structure of the visual scene from frame to frame across cuts (in film) and transitions (in interfaces).

## 2. THE PROBLEM OF WATCHING FILM

Our ability to perceive, let alone comprehend, motion picture films has long been recognized as a challenge by psychologists, especially for theorists of visual perception. Because of the optical distortions created by the camera, the image portrayed by a single frame of a film does not correspond to a conventional perspective view of a scene. When frames are projected to make a motion picture, and edited, the relation between normal vision and film becomes even more strained. The fact that we can still see objects and understand their behavior is informative because it means that the information provided by film does not violate the needs of basic perceptual processes.

As early as 1916, Hugo Münsterberg (1970) compared the close-up shot to perceptual attention, flashbacks to acts of memory and mental imagery, and the sequencing of shots to the sequential direction of attention around a real-world visual scene. Carroll (1980) reported that another early film theo-

rist, Pudovkin, described the role of the film editor as guiding the viewers' attention to certain elements of the scene, the laws of editing therefore being the same as those governing "ordinary looking." He and other analysts (Balázs, 1970; Eisenstein, 1949) also discussed the use of close-up shots to magnify critical details to the exclusion of the surrounding scene, in the same way that a viewer in the real world can concentrate on one part of the scene to the exclusion of the periphery of their gaze.

Mamet (1991) wrote the following:

> You always want to tell the story in cuts … if you listen to the way people tell stories, you will hear that they tell them cinematically. They jump from one thing to the next, and the story is moved along by the juxtaposition of images—which is to say, by the *cut.* (p. 2)

Lindgren (1963) compared film editing to prose narratives that describe a scene object by object, detail by detail, but where the objects are spread throughout a scene rather than being described in a linear spatial order. Lindgren noted,

> The fundamental psychological justification of editing as a method for representing the physical world around us lies in the fact that it reproduces this mental process in which one visual image follows another as our attention is drawn to this point and to that in our surroundings. (p. 62)

Although edited film, with its potential for large spatial and temporal jumps, can present scenes and sequences that are very different to those experienced in our real lives, it nonetheless relies on our ability to integrate different viewpoints and attentional foci into a single train of thought. This forms the basis for our cognitive analysis of film watching.

The value of applying an analysis of film to HCI was recognized by Hochberg (1986), who advocated the study of film techniques to aid the then emerging technology of computer-generated imagery. His argument was that, despite the gross differences between real-life scenes and the images contained in films, and the optical distortions created by the camera, filmmakers at least had the advantage of being able to point their cameras at real-world events; so many of the constraints on object construction, appearance, and behavior that our visual systems might make use of were implicitly recorded in the resulting film. Computer-generated imagery, on the other hand, had no such constraints; its scenes could portray anything, behaving in any fashion, at any level of veridicality, ranging from pixelated monochromatic wire-frame sketches to high-resolution, anti-aliased photographic renderings complete with multiple light sources, reflections, and receding surface textures.

In the two decades since Hochberg's (1986) work, the computer-generated image has become a ubiquitous component of commercial film, with the evolution of novel representational techniques such as "bullet time" (in which action slows to a crawl while the viewpoint revolves around a single element of a scene, such as a bullet speeding from a gun toward its target). These forms of portrayal have only become possible by digitally modifying film shot with many cameras and editing it together into a seamless whole. It certainly bears no relation to anything ever experienced by a human viewer of real events, and yet is comprehended instantly, on first sight, by every moviegoer. In itself, this is evidence against the commonly held view that film techniques are a form of convention or grammar that has to be acquired, and that film audiences can only see and understand film because they are immersed in a culture pervaded by film.

Filmmakers have learned through one century of experimentation what forms of dynamic scenes are easily comprehended by their viewers and which are not. Their craft knowledge delineates the comprehensible from the incomprehensible as well as any other body of empirical research that has taken a century to collect. The knowledge is embodied in many textbooks and handbooks written for trainee filmmakers, such as Maltby (1995), Richards (1992), Katz (1991), and Mamet (1991). Potentially, these books should provide a source for us to find principles that could be applied to HCI. Of particular interest to us is the way that different shots can be cut together, for in these situations the whole view portrayed on the screen changes, and yet viewers can quite easily make sense of a sequence of shots and may not even notice the cuts. However, these books are all situated within the domain of film, and it is not at all clear how the knowledge that they contain can be transferred to interface design.

Katz (1991), for example, concentrated on camera positioning and composition within a shot, rather than on the relative composition of shots; he confined his discussion of editing to the problem of when in the action cuts should take place. In HCI terms, this might correspond to the conventional concerns of what should go in a window, and where, and when it should open and close; it does not inform us about the spatial and thematic relations between different windows or different sequential views within a window. Richards (1992) also wrote mainly about composition within the shot, but did discuss "matching" and commented,

> If the subject is established in the right section of the frame, she must remain in that area even when you are cutting to another angle. When it is a reverse shot, logically one might think the placement of the subject on the reverse area is necessary. However, this is not the case. In fact it tends to confuse the audience. The simple theory is that the shifting of the viewers' eyes from one area to another

> confuses them … our acceptance of the cuts results from keeping the figure in the same frame area. (pp. 72–73)

This encapsulates the practical consequences of our argument about collocation and with practical implications for framing sequential shots.

Due to the experiential way in which craft knowledge is acquired by filmmakers, it is difficult to know when to apply particular principles, or to justify using one rule rather than another. They are all part of the craft of film that has to be learned by doing, by practical application of the apparatus of filmmaking. Much of the filmmaking advice is directed rightly toward rapidly and succinctly conveying narrative information or on leading the viewer to infer motive and intention, aspects of filmmaking that are not directly relevant to interface design. This makes it difficult to extrapolate the comprehension of dynamic scenes in general, and to interface design in particular. Even when a specific and relevant principle can be elucidated, such as the use of collocation in match cuts, it is not obvious when, or how, collocation should or should not be used in interface design. Although insightful interface designers who are also expert in film editing may be able to identify correspondences between particular editing techniques and a concrete design problem, a case-by-case approach to transferring knowledge between domains lacks generalizability, and justification (necessary to convince other designers) is slow and haphazard and prone to becoming rapidly outdated by advances in interface and device technologies. A more principled way of mapping the knowledge of film into the domain of interface design (and guidance on its use) is necessary to provide a rationale for each design recommendation.

May and Barnard (1995) argued that to transfer the craft knowledge from cinematography to interface design economically, an intermediate psychological account was needed, which described why the craft knowledge worked in terms of their consequences for the viewers' mental processing. By explaining why certain forms of film cutting work and others do not, in terms of the viewer's information processing resources, we seek to show that it is possible to derive principles that can be applied to the perception of dynamically changing scenes in general. By expressing these principles in a form compatible with wider psychological knowledge, we can go beyond the source material to make recommendations that are applicable to interactive dynamic displays in particular. In this article we elaborate May and Barnard's (1995) analysis; provide some empirical support for the assumption that filmmakers can and do use cutting to manipulate their viewers' gaze direction; and, using the theoretical analysis of the craft rules that have evolved in cinematography, suggest some guidelines for designers to follow to make dynamic transitions in their interfaces more film-like and, we argue, more comprehensible and usable.

## 3. FILM THEORY FROM FILM CRAFT

The most influential of the early film theorists was Eisenstein (Glenney & Taylor, 1991; Taylor, 1988). His "theory of montage" was an analysis of (initially) five types of montage that simultaneously coexist in any film sequence:

| | |
|---|---|
| Metric | Temporal length of individual shots. |
| Rhythmic | The relation between the temporal lengths of successive shots. |
| Tonal | Commonalities in attributes of objects in successive shots. |
| Overtonal | A sense or feeling emergent from the preceding three types. |
| Intellectual | A rational abstraction of meaning from a sequence of shots. |

This typology was largely developed to cover the patterns of cutting used in his silent, monochrome films, and he later added a sixth type (chromophonic montage) to deal with the synchronization of music and color. Each of these types attempts to isolate a particular form of interpretation or emotion aroused in the viewer by the sequence of shots. The typology seems to prefigure later semiotic approaches to film analysis (e.g., Metz, 1974) dealing as they do with the meaning that is to be inferred from the signs and conventions of film, rather than from the content of the shot.

Another early theorist who examined film was Münsterberg (1970), who related the sequencing of shots to the sequential direction of attention around a real-world visual scene. In essence, this is the line that we are taking in our assumption that the comprehension of film can be understood by recruiting cognitive theory. Taking Münsterberg's view with Eisenstein's (1949/1972), it is clear that the typology of montage must be comparable in some way to the perception of noncinematic visual scenes. The scene portrayed by the camera represents the standpoint of the viewer, and the assemblage of shots from a single scene should therefore be consistent with the views that a person might see if they were really there and able to direct their attention around the scene, despite the fact that the physical behavior of the objects and features represented in the film image is not the same as is experienced in real-world perception, as detailed by Hochberg (1986). The similarities must be at a more abstract level.

The first two types of montage are defined in terms of dynamic changes to the scene. *Metric montage* corresponds to the length of time the viewer spends looking at one particular point before turning their gaze to another focus (although in film, that focus may be 1,000 miles or years away). *Rhythmic* montage represents the frequency with which a viewer looks back and forth between two points (e.g., between two people conversing or between two intercut scenes, which may be separated as in metric montage) and hence

conveys a sense of the speed or pace of the interaction that is being observed. Both of these are critically dependent on the temporal dynamics of the editing. The other three are concerned more with the content of the shots being edited together.

*Tonal* montage corresponds to a highly generic perceptual ability that allows us to detect "common cause" in the motion or appearance of objects—Carroll (1980) cited Eisenstein as giving an example of the "dawn mists" sequence in *Battleship Potemkin,* which repeats a rocking movement in the motion of the water, the ships, the sea buoys, the sea birds, and the rising of the fog. Objects that share attributes as we look around us may also reveal some unobserved force or object. Similarly, overtonal, *intellectual,* and *chromophonic* montage all relate to abstractions that must be inferred from the content of the scene to understand the narrative of the film (or the nature of the situation in which a real-world viewer finds themself). In all of these, the identities, meanings, and associations of the objects being portrayed are more important than their physical characteristics or visual features.

A notable difference between the representation on the screen changing before a viewer's stationary gaze and the viewer having to move their head in a stationary world is the absence of proprioceptive feedback about the motion and new position of the viewer's head and eyes. This could mean that the viewer is not able to relate the new shot to the preceding viewpoint, and this is where the rules governing "allowable" cuts are generally invoked. A concise list of shots is listed by Bernstein (1988, pp. 160–167): establishing shot, close-up, reaction shot, cutaway, eyeline, eyeline match, jump cut, manipulation of time, and parallel action. Although these largely define shots or sequences of shots, rather than cuts, these all make sense if considered in terms of Münsterberg's (1970) vicarious viewer. They are, in general, aimed at ensuring that the points of view used in successive shots are consistent with those of a single observer moving their gaze and focus of attention through a scene, albeit with gross exaggeration and distortion of the scale of temporal and spatial change.

Hochberg (1986) also listed the "kinds and uses of abrupt transitions" in the visual scene that cuts caused, pointing out that it was not until the 1950s that "Hollywood developed the art of invisible, or seamless, cutting … the aim was to conceal from the viewer that a cut had been made" (p. 55). This was not a trivial problem, for many abrupt transitions between shots are readily perceptible and distract the viewer from understanding the scene. Kraft (1986) contrasted the role of *rhetorical* and *syntactic* cutting in films, where the former "influenc[es] the connotative and affective characteristics of film sequences" and the latter served as visual punctuation "segment[ing] the flow of filmed activities" (p. 155) to separate activities that needed to be parsed in the grammatical manner suggested by Carroll and Bever (1976). His

findings were that cutting did not serve a syntactic function, and that the number of cuts was not remembered. When viewers were asked to count the cuts, they were able to, but their recall for the activities portrayed was poorer. Therefore, they were able to attend to the physical structure of the film cutting, or to the meaning of the scenes portrayed, but not both (Kraft, 1986). Cuts that violated the cinematographic principles of "directional continuity" impaired viewers' ability to remember the underlying flow of action in the story, whereas varying the camera angle had no effect (Kraft, 1987). His conclusion was that "violating directional continuity disrupted viewers' expectations concerning cinematic space; these viewers were prevented from drawing the necessary inferences for representing the underlying actions" (Kraft, 1987, p. 11).

Of course, the unfolding narrative within a film provides a background constraint on what is being perceived; and if perception fails to make sense of the visual scene, the narrative can be recruited to aid comprehension. Nevertheless, for the most part, films do not present us with such perceptual challenges, and their visual slickness allows us to concentrate on the narrative without needing to struggle to make sense of the display, unlike many of our computing interfaces. A consensus between these theorists is apparent: Reconciling the abrupt visual changes that do occur in film relies on the same perceptual processes that allow us to make sense of the real world; unfilmic cuts present visual information in a way that is discordant with these processes, and hence do interfere with the narrative comprehension. To make a mapping from the language and technology of film to computer interface design, a model of these processes is an essential intermediate stage.

## 4. AN INTERACTING COGNITIVE SUBSYSTEMS MODEL

Our model is constructed within Barnard's Interacting Cognitive Subsystems (ICS) framework (e.g., Barnard & May, 1999; May & Barnard, 2003). The ICS framework represents human cognition as a sequence of transformations of information from incoming sensory representations, through a number of central mental representations, allowing the production of effector representations that control overt behavior (movement, speech, etc.). The transformations are grouped into subsystems that all deal with a particular form of representation. There are three subsystems dealing with *incoming sensory* representations (acoustic, visual [VIS], and body-state) and four with *central* representations (object [OBJ], morphonolexical [MPL], propositional [PROP], and implicational [IMPLIC]). Two further subsystems (articulatory and limb) transform *effector* representations into actions. For the purposes of this article, we need to consider the content of four of these levels of mental representation:

1. VIS: A sensory level of representation of the information extracted from the retinal image in terms of edges, features, hues, contrast boundaries, and so forth; unintegrated and prior to any organization of these features into shapes or objects. This level of representation can be transformed to produce OBJ and IMPLIC representations.
2. OBJ: A perceptual level of representation in which the visual scene has been parsed into coherent objects with orientation, spatial location, and depth including inferred physical characteristics not necessarily directly available from the visual scene. This level of representation can be transformed to produce PROP and limb (i.e., motor) representations.
3. PROP: A semantic level of representation in which entities within a scene have distinct identities, properties, and relations with regard to one another. This level of representation can be transformed to produce OBJ and IMPLIC representations and MPL (sound-based) representations.
4. IMPLIC: A holistic level of representation in which the propositional relations and sensory features of a scene combine to produce inferences about the real meaning or importance of a scene (i.e., what the entities are doing and why), drawn from the individual's experience of the world. This level of representation can be transformed to produce propositional representations and to create somatic (SOM) and visceral (VISC) changes in the body.

Each of the subsystems is able to receive representations in its own specific format, store them, and transform them into a limited number of other representations, as noted in the earlier descriptions. The cognitive models produced in ICS contain four main considerations. First, the behavior of the complete mechanism depends on the particular transformation processes that are required to support a cognitive task. This is referred to as a configuration of processes. As described later, the basic configuration for the interpretation of a visual scene requires visual input to be transformed into an object representation. This will form the basis for the derivation of a propositional representation of the events in the scene from which an implicational representation about their meaning can be inferred. There may also be subsidiary transformations that are of interest, such as that using the propositional representation to produce a mental verbalization (at the MPL level), but these are not part of the basic configuration. The basic configuration is thus written as VIS→OBJ, OBJ→PROP, and PROP→IMPLIC, although as seen later, additional feedback processes are active too.

Second, each process in a configuration is constrained by the recodings it has "learned" and can perform more or less automatically. This is referred to

as the procedural knowledge embodied in the process. If a process cannot easily transform a particular pattern of incoming information, then this will affect the overall operation of the complete system. For the interpretation of a film, we can assume that the procedural knowledge is all in place: that is, the images on the screen, the events portrayed, and their meanings will all be within the viewer's body of experience; the processes would, in principle, have little difficulty carrying out transformations of what is shown. This may not be the case in the comprehension of computer interfaces where the display elements are more likely to be symbolic and abstract, and so would be a clear point of departure for the modeling of cognition.

However, not all knowledge is represented in a proceduralized form. Some is held as a form of episodic memory. Within ICS, this is dealt with by the "image records" of the different subsystems. These are records of all representations that a subsystem has processed in the past and provide for a degree of abstraction over experience. Performance on a task may therefore also depend on the nature and properties of the memory records that need to be accessed. Therefore, the third consideration relates to the record contents accessed or used in the task setting. Of most relevance are the records that have only recently been laid down, and integration of these over the short term will allow for the recognition of just-seen objects, events, and so on.

These considerations deal with the capabilities of individual subsystems. Because the dynamic course of cognition depends on interactions between subsystems, a fourth consideration deals with properties relating to the overall dynamic coordination and control of the mechanism. This is referred to as *dynamic control* and is intended to capture the status of representational resources, the extent of their use, and how information flow is coordinated and used in the internal monitoring and evaluation of configural activity. Because the individual subsystems are independent and act in parallel, this consideration is emergent from the requirements of the flow of information through the mechanism as a whole—it should not be thought of as implying the existence of some "central executive" that sets up a configuration, controls memory access, or schedules competing tasks.

For our purposes, dynamic control can be thought of as a "bottleneck" in processing: For example, if there is a problem with the transformation of the propositional representation, then additional processing will be required in the form of exchanges between this subsystem and others to elaborate and refine the content of its input. The limits of the propositional output will constrain the performance of subsequent processes that operate on it, and so the PROP subsystem will become the locus of dynamic control. Dynamic control can shift between subsystems according to the task. If people are asked to notice each occurrence of a particular object, then they will have to use the object representation, because this is where that information is held.

As a framework, the ICS perspective holds that an understanding of the cognitive underpinnings of behavior in complex tasks can be achieved by specifying properties of configurations, procedural knowledge, record contents, and dynamic control. By characterizing the information that is available for each process to operate on, we must make assumptions about the record contents and the degree of proceduralization of each process. An ICS model takes such approximate estimates as starting points and builds on them to map out the subsequent course of cognitive activity.

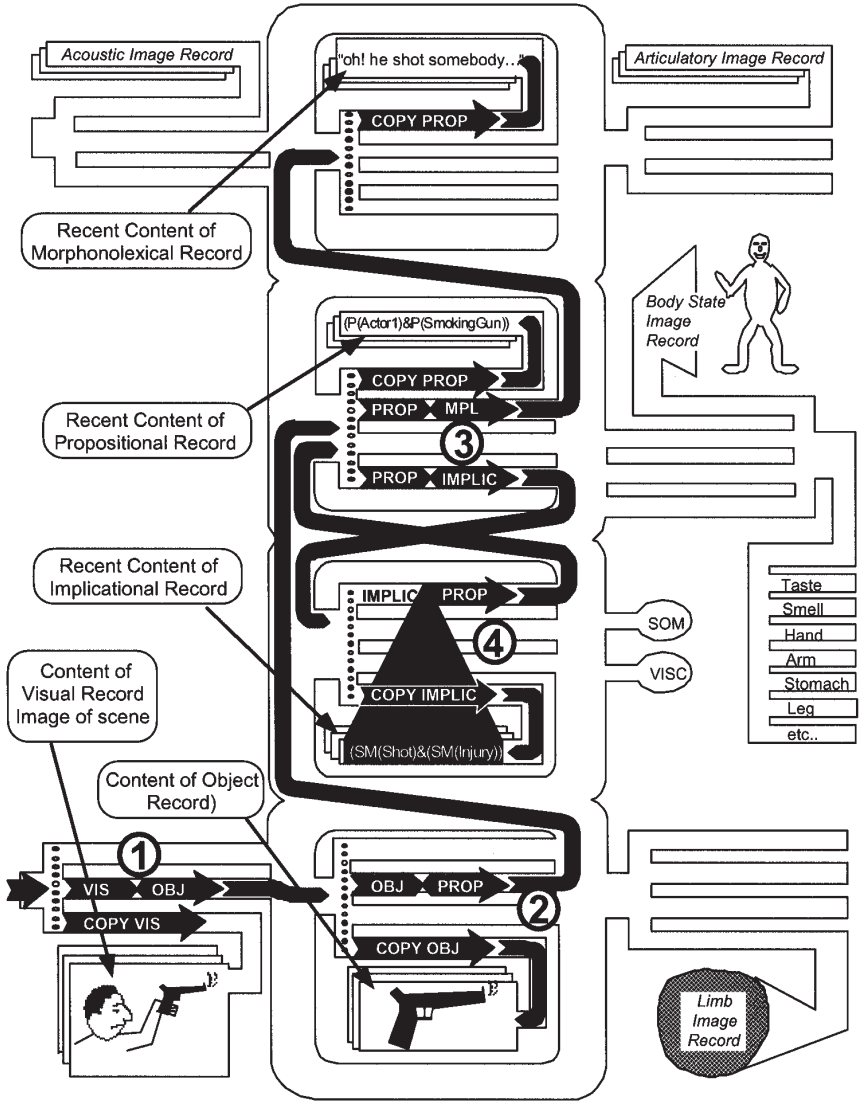
## 5. MODELING FILM WATCHING IN ICS

In this section, we outline the set of processes in ICS that are involved in the perceptual recognition of events portrayed in a film sequence and their assimilation with the narrative. We restrict this model to those aspects that are relevant to the problem of comprehending dynamic visual scenes, and so do not attempt to include in our model the more abstract questions of the comprehension of the narrative itself. In principle, this would be the work of the propositional and implicational levels of representation and, as we indicate, the viewer's understanding of the narrative is a component of their comprehension of the scene; but for now we can treat it as a given, without trying to model it explicitly. The reason for this limitation is that, as indicated in Section 1, the comprehension of a narrative is less important in HCI than in film, and we wish to give guidance that does not require interface designers to deal with these issues.

The configuration of processes illustrated in Figure 1 shows those that would typically be required for a viewer of a film to observe and comprehend the events portrayed—for simplicity, we deal with the visual input only; but in principle, ICS can also be used to model the simultaneous acoustic input, as we indicate. To begin with, the scene itself must obviously be processed through the VIS subsystem (1). A copy process creates an image record of all the information in the visual field—in Figure 1, the viewer has just watched one of the actors raise and fire a gun. The full scene appears as the representation in the image record of the VIS subsystem. In parallel, the VIS→OBJ process recodes the raw visual data into a higher order object representation, which reflects a more abstract structural description of visual form. This process of recoding involves information reduction as well as elaboration. For example, the object code would not represent gradations of brightness, but would now distinguish the form of objects.

Because these processes can only recode a single data stream at a time, the actual recoding would also be selective. With the recoding of a visual representation, selectivity would operate spatially and only part of the visual field would be undergoing recoding (the "attended-to" element). Here the viewer

*Figure 1.* **The configuration of processing involved in the comprehension of a dynamic scene. VIS = visual subsystem; OBJ = object subsystem; PROP = propositional subsystem; MPL = morphonolexical subsystem; IMPLIC = implicational subsystem; SM = schematic model; SOM = somatic changes; VISC = visceral changes.**



may be attending to one part of the screen (e.g., the gun). The object level description of this element of the visual field is copied into the image record of the OBJ subsystem. Note that the information represented in the OBJ subsystem's image record is qualitatively different to that in the visual image record, reflecting the processing that has taken place.

Simultaneously, the OBJ→PROP transformation process in the OBJ subsystem produces a propositional representation of the information (2). In this recoding, the details of the appearance of the elements in the visual scene are discarded, and an abstract semantic representation is created. This is copied into the image record of the PROP subsystem—in this example, the propositional information linking the actor, P(Actor1), with the gun, P(SmokingGun).

The figure shows two transformation processes occurring in the PROP subsystem (3). One, PROP→MPL, produces a morphonolexical representation from the information represented propositionally. Subjectively, this code corresponds to what we *hear* as our internal mental voice or imagination, and is descriptive of the propositional representation; but the process is not central to the basic configuration that builds the viewer's understanding of the scene. At the same time that the PROP→MPL process is generating the internal speech, the PROP→IMPLIC process is also active (3). In this transformation, details of the individual propositions are lost and the highest level cognitive representation of the scene is constructed, the implicational representation. This process involves interrelating propositions, both among themselves and in relation to prior experience as represented in the propositional image record. This process is therefore inferential in nature and abstracts the overall meaning of the constituent propositions when taken as a whole. The resulting schematic model (SM) is copied into the implicational image record. The information encoded within it reflects the overall conceptual structure of the narrative. Here the viewer has inferred that a shot has been fired (SM(Shot)) and someone has been injured (SM(Injury)), although this has not yet been shown. This model, being implicational, is not restricted to the bald fact that someone has been hurt but includes subjective feelings of shock, surprise, and threat; its activation causes SOM and VISC changes within the body of the viewer, via the IMPLIC→SOM and IMPLIC→VISC processes (not shown in Figure 1).

The chain of cognition from the visual information in the scene, through the object and propositional representations, has thus resulted in the viewer generating an implicational understanding of the action portrayed, and an anticipation of its consequences. However, this is not the end of the chain. We can see that the IMPLIC→PROP process is active (4), generating a further set of propositional representations from the implicational knowledge. The output from this process may be combined with the other inputs to the PROP subsystem. In this example, the consequence is that the input to the PROP

subsystem is not just the bottom-up output from the OBJ subsystem's interpretation of the visual scene, but also the ongoing top-down effort of the viewer to put what they are seeing into context, to construct the narrative. Here the locus of dynamic control is at the IMPLIC subsystem, and the IMPLIC→PROP transformation is not taking its input directly from the information reaching the subsystem, but is operating in a buffered mode, transforming the data in the proximal region of the image record. Because the "copy" process is continually transferring the input to the image record, the effect of buffering is to free the IMPLIC→PROP transformation from the timing constraints imposed on it by the pace of the incoming data and to let it work at a speed appropriate to the formation of coherent propositional output. The representation that it returns to the PROP subsystem, and is blended with the output of the OBJ→PROP transformation, is therefore more likely to be complete in terms of its consistency with the viewer's SM of the scene they are viewing.

It is interesting to note the overall similarity between the content of the various levels of representation and Eisenstein's (1949) forms of montage: the visual level covering the metric and rhythmic changes, the object level covering the tonal, the PROP covering the intellectual, and the IMPLIC covering the overtonal. In its distinctions between different levels of mental representation, the ICS framework inherently captures this early, film-based typology. It is the different qualitative natures of these levels that give rise to the mental phenomena associated with watching and comprehending an edited film.

To summarize, in this ICS model the viewer is focally aware and concentrating on the narrative meaning of the film, represented at the implicational level. Reciprocal activity between the propositional and implicational levels attempts to interpret new information from the scene in terms of this narrative. The VIS and OBJ subsystems extract information from the seen images, with the OBJ subsystem interpreting entities in the visual scene in conjunction with feedback from the PROP subsystem about what is likely, given the current understanding of the narrative. Comprehension of the visual scene is very much subsidiary to comprehension of the narrative. In what follows, we argue that inconsistencies between the subsidiary task and the primary task distract the viewer from comprehending the narrative, and so are generally to be avoided unless such an interruption is desired, perhaps for rhetorical effect.

If a cut were to occur at the point in the film represented in Figure 1, the new visual scene presented to the VIS subsystem would result in a different visual structure being transformed into object code. The OBJ→PROP process would in turn generate a new propositional representation, and this would have to be blended with the propositional output of the IMPLIC subsystem, which of course is based on the preceding scene. Broadly speaking, a cut will be acceptable if the information extracted by the viewer is consistent

with the representations currently active in the configuration of processes. Any inconsistencies or ambiguities could shift the locus of dynamic control away from the IMPLIC subsystem in an attempt to resolve the difficulties caused by the cut. Because the interpretation of the narrative must be carried out by the IMPLIC subsystem, the viewer would lose track of what is going on. For clear comprehension, then, a dynamic display or a film should avoid locating dynamic control away from the IMPLIC subsystem.

The whole point of cutting film is to move the viewer rapidly through a narrative sequence, without waiting for the camera to pan from side to side within a scene, or to move from place to place to follow the action. Cutting makes it possible to successively present views that are not spatially or temporally connected. For the resulting sequences to be comprehensible, there must be some connection between successive shots to enable the viewer to relate them to each other and to the narrative. This relation can be based on the object representation—when there are similarities in the abstract visual structure displayed on the screen, or it can be based on the propositional representation—when there are similarities in the actual elements displayed and their relations.

There are two points at which difficulties in comprehending a cut might occur. The most obvious would be an inconsistency of the novel propositional representation with what has gone before (i.e., an unexpected entity appears on screen, or entities appear in an unexpected relation or position). This would prevent the viewer from relating the new scene to their expectations of the narrative, derived from implicational representations. Problems could also result from cuts that make it difficult for the VIS subsystem to produce an appropriate object representation (i.e., the attended object's visual attributes change substantially). Because these could lead to propositional difficulties, we examine the object representations first and then turn to propositional representations. However, before doing so, it is worth restating the importance of implicational representations in the comprehension of the narrative.

Implicational representations provide a basis for the inferential processing needed for the viewer to construct the narrative as the events unfold before them, and we have mentioned their role in providing affective tone to the experience (a more detailed account of the role of implicational representations in affect can be found in Teasdale & Barnard, 1993). In terms of narrative comprehension, these are the most important representations because if these cannot be formed adequately, the viewer would be able to report nothing more than the appearance and disappearance of the objects and actors (from their object representation), or the sequence of events (from the propositional representation). Only the implicational representation contains the SMs that correspond to the viewer knowing why the events happen and what they mean. For those who seek to understand film, per se, the focus has rightly

been on this level of meaning; however, for our current purposes it is appropriate to focus on the representations supporting the perception of the visual scene: the object and propositional levels of meaning.

## 5.1. The Role of Object Representations

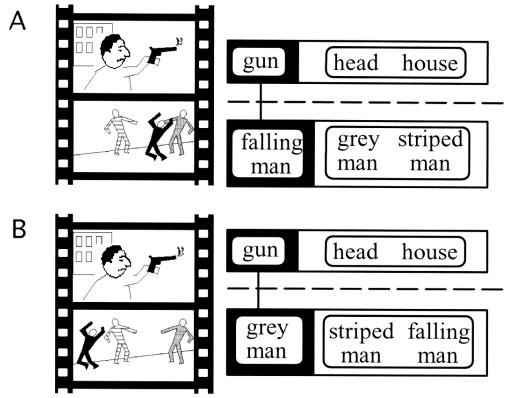
No theory of dynamic scene perception would be applicable without a technique or method for analyzing the temporal changes in a viewer's representation of a scene. To support the modeling required by the ICS theory, we have developed transition path diagrams (TPDs), a notational technique that enables designers to detail the thematic transitions in the topic of processing, step by step, as the scene changes and as the viewer's attentional focus moves. The rows of the notation correspond to successive attentional fixations on a processing topic (the psychological subject of the representation, shown in a black frame), with the superordinate grouping of that topic, its predicate, and its constituent structures detailed. Our central argument is that cinematic cuts that follow normal thematic patterns will be transparent to the viewer, and they may not even notice their occurrence. Cuts that do not replicate the effect of a thematic transition, however, will not feel natural and may be noticed.

An example of a common cinematic practice that illustrates this are match cuts, which place the element of the new scene that the viewer should attend to in roughly the same physical location on the screen as the probable psychological subject of the previous shot. In the example illustrated in Figure 2, after the actor has raised and fired the gun it is highly likely that the viewer will have been following the motion of the gun. If the succeeding cut to the gunman's target placed it in roughly the same screen location as the gun, then it would immediately form the psychological subject of the viewer's object representation. Placing it elsewhere would require the viewer to make a transition out to the superstructure of the scene, and then in again to a possible target before they could understand what or who had been shot at.

When there is a cut in a scene, the view changes such that the psychological subject of the object representation suddenly disappears. The VIS subsystem must take the new incoming information and transform it into a representation for the OBJ subsystem. If the visual structure of the new shot is such that there is nothing in the location of the previous subject that can be used to form the subject of a new representation, as in sequence B of Figure 2, the VIS subsystem will take longer to form a new object representation. The OBJ subsystem, meanwhile, will be without a meaningful input and will continue to base its outputs on the last input it did receive.

When the VIS subsystem has succeeded in reorienting itself to produce an object code output, the new information will replace that being operated on

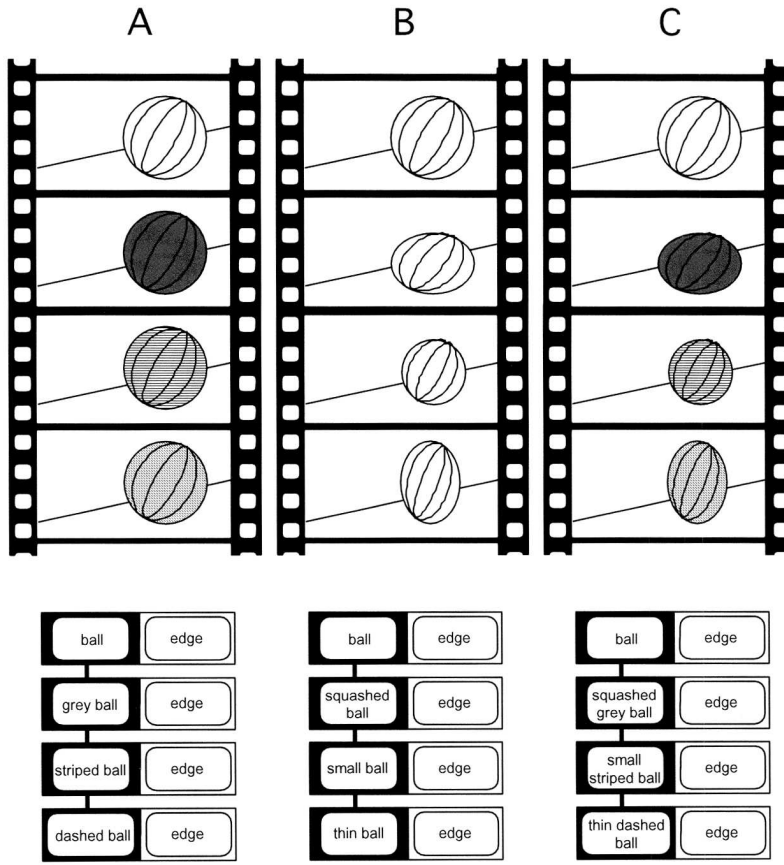
***Figure 2.*** **The cut in sequence A better conveys who has been shot than the cut in B, because in A the relevant element of the second shot (the falling man) is collocational with the subject of the first shot (the gun). A visual transition must be made after the cut in B to reorient the viewer's representation from the gray man on the right to the falling man on the left.**




by the OBJ subsystem. The smoothness with which this change can be dealt with depends on the similarity between the structures of the two representations. The greater the similarity between the old and new subjects (in terms of their visual attributes such as shape, color, motion, texture, etc.) and their respective predicate structures, the easier it will be for the transformation processes active in the OBJ subsystem to continue to produce output representations (i.e., propositional representations). Therefore, if a cut were made from the middle shot of an actor firing a gun to a close-up of the gun, or to a long shot of the whole scene, the VIS subsystem could retopicalize on the new position of the gun, although it may be much larger or smaller and in a different part of the screen, and the identity between the resulting object representations and that of the preceding middle shot would enable the OBJ→PROP transformation to proceed seamlessly.

Filmmakers can capitalize on this ability of the OBJ subsystem to cope with changes in the attributes of the subject of its representation. Cutting between two highly similar, collocated subjects could lead the viewer to interpret them as being representations of the same object, some of whose attributes have changed, rather than as representations of different objects (Figure 3). Of course, the more attributes that change, and the less spatially related the successive objects are, the less likely this is to succeed. A similar technique is extensively used in animation, where elements can undergo spectacular transformations of color and shape quite unlike those of their real-world counterparts without producing any propositional problems for the view-

***Figure 3.*** **The objects cut together in film A and in film B may be viewed as single objects whose surface design or shape changes, respectively (the cuts not being apparent), whereas the objects in film C will be viewed as different objects (the cuts will be apparent). The transition path diagrams below each strip show the relative complexity of changes occurring to the psychological subject in each case.**




ers—their reaction to the bizarre incongruities in the implicational representation being amusement.

As Figure 2 showed, collocation can be an important cue as to the element of the new shot that should be taken to form the subject of the representation, if it differs from the previous subject, or if there has been a gross change in some attribute (such as its shape, due to camera angle). In terms of a structural description of the scene, collocation maintains the relation between the subject and the psychological predicate of the object representation, at the possi-

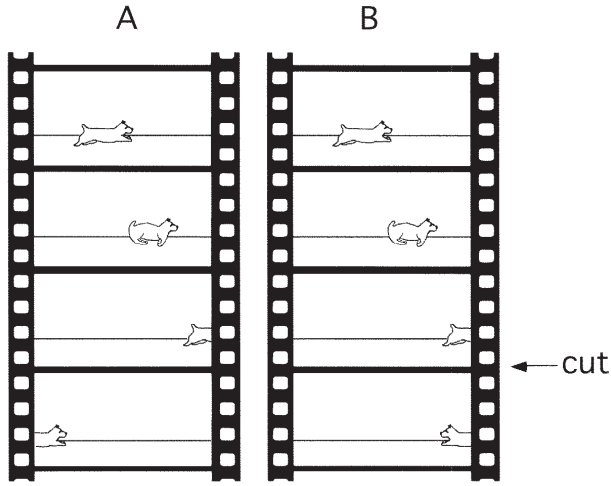
ble cost of changing the attributes of the subject and the elements in the predicate. Where the same element should form the subject in succeeding shots, however, it is usually more important to ensure that its attributes, rather than its location, are preserved. This may require translation of the subject to a new position on the screen, rather than collocation.

An example of this occurs when the element that is the subject of the representation is moving out of the frame of the shot. Then, if there is a cut to a new shot, which the element is to enter, preserving its location within the scene is entirely artificial because the cut is mimicking a change in the position of the viewers' head as they turn their gaze to follow the subject's motion (i.e., their point of view). Under these circumstances, the subject's position in relation to the field of view (i.e., the relation between it and the predicate structure) will necessarily change.

Maintaining the predicate structure of the representation requires the subject's attribute of motion to be changed, altering the relation between the two shots (Figure 4). Preserving the attribute of motion and allowing the predicate structure to vary creates a cut that is readily understandable, corresponding as it does to what would be experienced by a real-world observer moving their head. Sequence A in Figure 4 (modeled on scenes used by Frith & Robson, 1975) conveys the sense of the camera having turned to track the dog (even if the real position of the camera has moved); whereas in sequence B, where the camera has had to move to the other side of the dog's path, the impression is of the dog turning around. The camera remained stationary (in practice the background will have changed too; but even if this is sufficient to contradict the impression of a stationary camera, it will still not prevent the impression of the dog's change in direction—such a cut would be even more opaque than sequence B).

The collocation of the dog over the cut in sequence B is counterproductive because the VIS→OBJ transformation can proceed without needing to retopicalize to find a subject. There is no gross change in the predicate structure of the representation. The OBJ→PROP transformation is therefore likely to interpret any changes in the subject's attributes as real, rather than being due to some change in the point of view. As in sequences A and B of Figure 3, where a single ball appeared to persist with changing attributes, in sequence B of Figure 4 the dog appears to have turned around. This would result in a propositional representation quite different to that of sequence A, where the retopicalization by the VIS subsystem produces different object representations before and after the cut. In this sequence, the subject's attributes have been preserved but its predicate structure has altered, allowing the OBJ→PROP transformation to infer a change in the point of view. The elements being viewed continue their behavior and relations despite the cut.

***Figure 4.*** **When the dog exits right, a cut preserving its location within the structure of the scene (i.e., collocation) would require it to enter the succeeding shot from the right (B), altering its attribute of motion. Preserving the attribute of motion requires it to enter from the left (A), an example of translation. The impression gained by a viewer of the two sequences is quite different.**




These three examples of cutting in film have illustrated how a structural description of the object representations, produced by the VIS subsystem, can help explain the confidence with which certain cuts can be used. The TPDs help to embody the assumptions about the transitions that are needed, or which are most likely to occur, and difficulties in making these transitions result in additional dynamic control requirements within the configuration. The following summarizes the analyses so far:

- Keeping the psychological subject and its immediate predicate structure constant corresponds to cuts that close in or open up the shot (e.g., long shot to middle shot to close-up, and vice versa). These are the most readily comprehensible transformations of the visual scene, with no extra processing required by the VIS or OBJ subsystems.
- The subject is translated and its predicate structure changed (i.e., its location on the screen and its surroundings change); then, following extra processing by the VIS subsystem to relocate the subject, a change in the observer's point of view is assumed by the OBJ subsystem without disturbing its processing.
- When the predicate structure and the psychological subject both change (i.e., a cut to a different point of view, with a different element collocated

with the previous subject, even if the previous subject is still somewhere on screen), then the VIS subsystem will use the new element to produce the psychological subject of the object representation.

However, this last situation has consequences beyond the VIS→OBJ transformation, as do cuts where there is no collocated element, and where the previous subject cannot be found within the scene. Here the OBJ subsystem cannot maintain any continuity between the shots, and the PROP subsystem has to resolve their relation.

## 5.2. The Role of Propositional Representations

The VIS→OBJ transformations described in the previous section dealt with the identification of individual elements of the scene, before and after a cut. Watching a film, of course, involves more than following objects around from shot to shot. The second step is to understand their relations with each other, and this is carried out by the transformations from object to propositional representations (OBJ→PROP). Where the object representation is an abstract structural description of entities and relations in visual space, the propositional representation is a description of entities and relations in semantic space.

In the example configuration of Figure 1, the viewer has been looking at the actor and has watched him raise and fire the gun. The current subject of their object representation is the gun, from which smoke has just appeared. The OBJ→PROP transformation uses this object representation to recognize this object as a gun, to link the smoke spatially with the gun, and to identify the composite as a gun that has just fired. The element "Smoking Gun" is thus the psychological subject of the propositional representation, and "Fired, by Actor 1 … " the start of the predicate. This representation is used as the basis for the PROP→IMPLIC transformation that would give rise to the understanding that something else had just been shot by Actor 1, as well as producing the implicational feelings of shock, surprise, and threat, which serve to contextualize the subsequent processing and give film watching a sense of engagement that would be lacking if it were watched at a solely propositional level.

The cuts illustrated in Figure 2 would help to clarify just what had been shot. As explained in the previous section, sequence A of Figure 2, where a person throwing their arms up and collapsing is collocated with the gun, would lead to an object representation with this person as the psychological subject. The propositional representation of "Smoking Gun, Fired by Actor 1" would thus be succeeded by one with "Person (Surprised, Falling)" as the subject. The implication that the falling person had been shot by Actor 1 would be easy to draw.

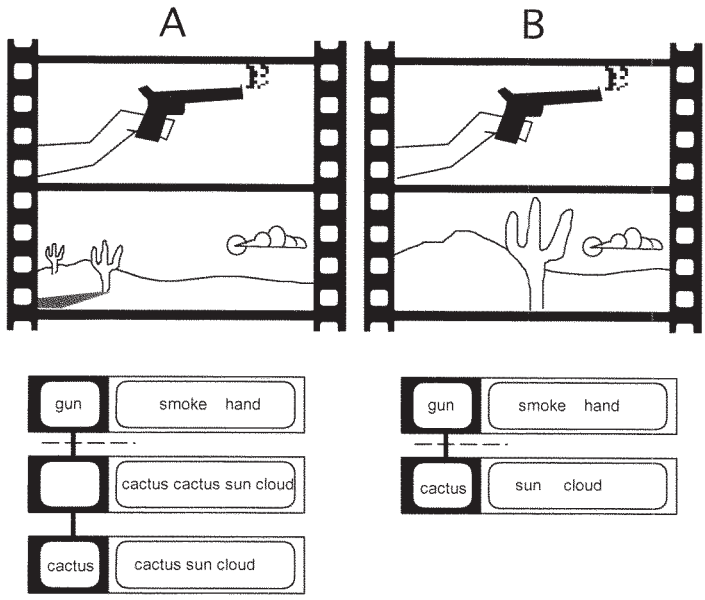
The implicational representation generated by the sight of the smoking gun would have been used by the IMPLIC→PROP transformation to feed back the proposition, "Something, shot by Actor 1." On the cut, this proposition can be combined with the information coming from the OBJ subsystem to produce the richer representation of "Person (Surprised, Falling), shot by Actor 1" immediately, without further processing. Instead of seeing a shot fired, then seeing someone fall, and "working out" that they have been shot, the viewer sees a shot fired and then sees someone being shot.

In sequence B, however, the VIS subsystem is likely to have retopicalized on the central figure, leading the OBJ→PROP transformation to produce a propositional representation of "Person (Leaning, Looking screen-right, next to Person [Surprised, Falling])." This cannot be successfully combined with the proposition arriving from the IMPLIC subsystem. One of them would predominate, with the viewer either focusing on the leaning person, then following their gaze to search the right-hand side of the screen, and so not noticing whom Actor 1 had shot—or focusing on the unsatisfying proposition that something has been shot and starting a propositionally driven search of the scene to find who or what. If the next cut followed too quickly, then they might not succeed in making the connection at all. In neither case does sequence B of Figure 2 convey whom Actor 1 shot as clearly as sequence A.

This shows how cuts that force the viewer to form novel subjects in their object representation rely on the PROP subsystem to make sense of the scene. The interplay of OBJ→PROP and IMPLIC→PROP will also be active when there are no intended propositional links between the subjects of succeeding shots. When one scene ends and there is a cut to a completely new scene, the filmmaker must take care that the viewers do not interpret the new view in the light of the preceding shot. A safe way to do this would be to do what is not recommended earlier. The filmmaker should ensure that there is no collocated subject in the second shot (Figure 5), and that the subject of the first shot does not recur in the second shot; or if it does, then it should have quite different attributes.

The tendency of viewers to carry over propositional and implicational information from one shot to the next can be used for narrative effect, of course. If a collocated subject is present in the second shot, then the propositional attributes of the preceding subject may be blended with it, providing an additional level of allusion not available from a single shot. This is what is meant by Eisenstein's (1949) "overtonal montage," and a good example can be seen in Kubrick's (1968) film *2001: A Space Odyssey*. In the first scene—an allusion to the biblical story of Cain and Abel—a primate who has been cognitively enhanced by the monolithic Sentinel has learned to use a thigh bone as a weapon and has just slain a member of another tribe to gain access to a water hole. At the end of the scene, the primate hurls the bone upward, and the

***Figure 5.*** **At the end of the first scene in sequence A, the cut to a long shot with no collocational subject makes it less likely that viewers will make any propositional links between the shots. The middle shot with the collocational cactus in sequence B could mislead viewers to associate the cactus with the consequences of the gunshot.**




camera tracks it as it spins against an empty blue sky. Then, there is a cut to a space station drifting in space.

Here, Kubrick (1968) presumably wants the viewer to make an implicit association between the warlike nature of the thigh bone, made into a weapon by intelligent action, and the militaristic nature of the space station. He does this by using the principles outlined in the previous section—the bone and the space station are in roughly the same location of the screen, they share several attributes (both are long, thin, and white), and they have a common predicate structure (moving against the sky). He further removes the chance of the viewer focusing on an irrelevant element before or after the cut by making the bone and the space station the only elements in their respective shots. On the cut, the viewer's object representations are therefore highly similar, there are no interruptions to the flow of processing from the VIS subsystem through the OBJ subsystem to the PROP subsystem, and there is little to challenge the blending of the propositional representation derived from the implications of the previous shot blending with the new subject of the propositional representation. Indeed, if the background sky did not change from white to black,

viewers might not realize that the cut divided two scenes, and so might briefly think that the bone had actually changed into a space station.

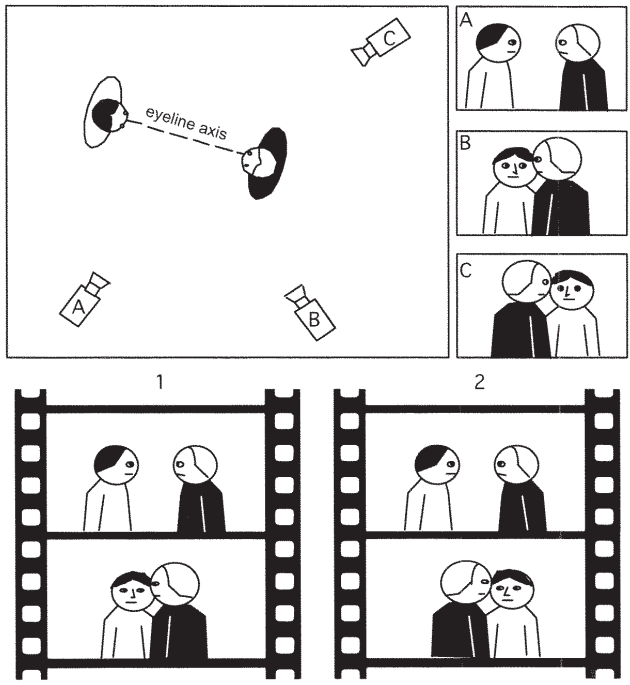
Although these end-of-scene cuts are important, cuts that result in a novel visual structure are more frequent within a single scene, corresponding to different views of various elements as the action continues. Unlike the Kubrick (1968) cut, these more common internal cuts are generally designed not to be noticeable. In these cases, the PROP→IMPLIC and IMPLIC→PROP cycle should be allowed to continue as smoothly as possible, to avoid distracting the viewer from the narrative. If the change in the predicate structure and the attributes of the subject are such that the OBJ subsystem can infer a movement of the observer's point of view, such as change in position or a turning of the gaze, then this subjective motion will be included as part of the propositional representation, but only as part of the predicate structure. This will not interfere with the implicational cycle and may be effective in giving the viewers a sense of their own involvement in the scene, as they apparently move with the actors (sequence 1 in Figure 6).

If the change is too gross, however, the OBJ subsystem may infer a jump in the absolute positions of some elements of the scene, leading the viewer to attempt a comprehension of the jump as part of the narrative (sequence 2 in Figure 6). An even grosser change may prevent both retopicalization by the VIS subsystem and an inference of subjective motion by the OBJ subsystem. This would lead to a complete dissociation of the propositional representations, and the viewer would be likely to interpret the new shot as an unrelated scene. In these two situations, subsequent IMPLIC→PROP feedback would be hard to assimilate and may soon indicate that they have misinterpreted the cut. They would then have to reassess their evaluation of the scene, possibly by accessing the image record of their PROP subsystem rather than by attending to the incoming propositions from the VIS→OBJ, OBJ→PROP path.

The cut shown in sequence 2 of Figure 6 is an example of crossing the axis, where the two camera positions are on alternate sides of an imaginary line drawn between the two actors. This is generally accepted as bad practice and makes certain situations very difficult to set up for filming—Boorstin (1991) warned about dinner table scenes, where for $n$ people there are $n(n-1)/2$ different axes that must not be crossed. He described Woody Allen's (1986) solution to this in *Hannah and Her Sisters* was simply to avoid cutting and to slowly rotate the camera around the periphery of the table, letting each character talk as they came into the shot (conceptually simple, but demanding on the actors).

Even when there are only two characters, remaining on one side of the eyeline axis can cause problems. Boorstin (1991) described the problems involved in filming scenes where two characters converse in the front seats of a moving car (Figure 7). If the eyeline axis is not crossed, then the predicate structure of the shot must, due to the confined space available for setting up

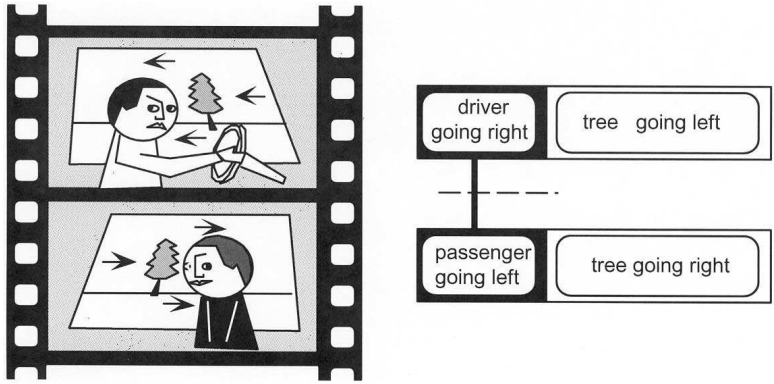
***Figure 6.*** **Cutting from camera A to camera B (sequence 1) is acceptable because whichever actor is the viewer's psychological subject, there is collocation following the cut, and the predicate structure is maintained. Cutting from A to C (sequence 2), however, neither presents collocation or maintains the predicate structure of the scene and would not be acceptable.**




the shot, be of the landscape visible through the window behind them. For the character in the left-hand seat, the landscape will be moving from right to left across the screen, but when the cut is made to the character in the right-hand seat, the landscape will be seen to move from left to right. The uncomfortable consequence of this sort of sequence can, in our analysis, be explained as a gross disparity in the predicate structure of the scene—the two characters will appear to be in two different cars going in different directions. Although crossing the axis can be overcome quickly by the top-down constraints of the narrative, it momentarily distracts the viewer from the message of the film and makes them attend to the medium. As with all violations of the stylistic rules cited here, this can have benefits when used sparingly, but becomes tedious if overused.

This section described the consequences for the viewer's propositional representations of various forms of cut that can be used in film editing. Where

***Figure 7.*** **Because of the different flow of elements in the predicate structure of the two shots, the characters in the front seats of a car appear to be traveling in different directions.**




the OBJ subsystem could deal with a range of cuts affecting the abstract visual structure of the scene, the PROP subsystem can make inferences over cuts that manipulate the propositional representation in certain ways. The key points are as follows:

- Viewers will tend to make implicational inferences between successive propositional subjects. This can be valuable if used carefully, but can be suppressed by avoiding collocation or translation of subjects over the cut.
- Changes in the predicate structure of an object representation that cannot be reconciled with a change in the viewer's position or gaze will be represented propositionally as a change in position of the elements of the scene.
- Shots that cannot be propositionally linked will be interpreted as unrelated, unless the implicational representation of the narrative makes it apparent that there was a relation.
- In the absence of any support for propositional retopicalization, viewers will carry out exchanges between the central subsystems to resolve the ambiguity, neglecting the ongoing action.

## 6. DO FILMMAKERS REALLY MANIPULATE GAZE DIRECTION?

An untested assumption of the entire argument so far is that filmmakers do actually attempt to manipulate gaze direction and that their attempts succeed.

Although the language of film editing includes terms such as *match cut* and *jump cut*, this does not necessarily mean that filmmaking practice routinely follows the prescriptive advice of the textbooks. Even if cuts do use collocation appropriately, there is no guarantee that viewers' gaze direction is guided toward the appropriate objects. These two assumptions need to be confirmed if this particular mapping from film to computer interfaces is to have any value. To verify these assumptions, we decided to collect some empirical evidence about the gaze direction of the viewers of a commercial film. Our intention was twofold: first, to find out whether we could detect changes in gaze direction contingent on a cut in the film; second, to see if we could distinguish between cuts that did result in changes in gaze location from those that did not, on the basis of our theoretical analysis.

The approach that we took to test this was to record the eye movements of some volunteers while they watched a full-length commercial feature film, in its full theatrical aspect ration (in contrast to the "pan and scan" version made for TV broadcast or VCR distribution). By measuring the location of gaze directions before and after each cut in the film, we can find out whether our volunteers tended to keep their gaze fixed at the same point on the screen following a cut, or whether they tended to look at a different place shortly after the cut. The theoretical analysis that we have given so far predicts that some types of cuts should make use of collocation, and so should not result in a change in gaze direction, whereas other types should not use collocation, and should result in changes in gaze direction. By looking at the relative positions of gaze direction around a cut, instead of the absolute position, we avoid the need to define specific regions of interest for each shot, which would involve guessing the filmmakers' intentions. In effect, we are just looking at the consequences of the editing process. This is important because if we (as experimenters) were to attempt to identify the objects that we (as viewers) felt the director wanted us to be looking at before and after the cut, and then found that our volunteers also looked at the same objects, we would have found out no more than that they were watching the film in the same way that we were. Measuring only changes in relative gaze direction is not only simpler, it is more objective.

We took a DVD version of the Columbia Pictures film *The Mask of Zorro* (Campbell, 1998) and identified every cut. In the film's 132 min, there were 1,916 cuts (roughly one every 4 sec). We then classified the cuts according to an objective taxonomy, based on what is actually portrayed on the screen before and after a cut, rather than being expressed in intentional terms (such as Bernstein's, 1988, set of nine sequences listed earlier).

The classes of cuts are listed in Figure 8, together with their frequency of occurrence in the film. Seven of the classes, which are described later, were defined a priori, but three were added during the classification process (specifically, over-the-shoulder, previous, and topical cuts) as special instances of

***Figure 8.*** **Classification of film cuts, together with number of each within the film *The Mask of Zorro* (Campbell, 1998), predicted benefit of collocation, and time of statistical peak in eye movements.**

| Class | Description of Shots Before and After Cut | *n* | Benefit From Collocation | Peak Eye Movement (msec) | *F*(1, 45) |
|---|---|---|---|---|---|
| Detail | Closing in or opening out from a detail | 54 | Yes | No peak | *ns* |
| Result | Result of some cause or action shown in first shot | 61 | Yes | 280 to 440 | 8.15** |
| Following | Different views of moving person or object | 276 | Yes | No peak | *ns* |
| Conversation | Shots of two or more actors involved in a conversation | 560 | No | 120 to 200 | 37.5*** |
| Subjective | First shot of an actor looking at object; second shot of that object without actor in shot | 142 | No | 120 to 280 | 18.8*** |
| Over the shoulder | First shot of an actor looking at object; second shot of that object with rear view of actor in shot | 26 | Yes | 120 to 280 | 7.03* |
| Previous | Second shot returns to view used previously with no more than two intervening shots | 149 | No | 120 to 200 | 4.70* |
| Topical | Second shot contains objects whose presence is predictable from first shot | 341 | Yes | 120 to 280 | 46.3*** |
| Novel | Second shot contains new scene or objects whose presence is not predictable from first shot | 95 | No | No peak | *ns* |
| End of scene | Unrelated in time or place; start of a new scene | 27 | No | 200 to 280 | 10.1** |

*Note.* Ratios were nonsignificant at the $p > .05$ level.
*$p < .05$.**$p < .01$.***$p < .001$.

other classes. An initial classification was carried out by Jon May, and then reexamined by Michael P. Dean, with 185 cases (9.7%) being queried. A consensus reclassification was reached in each case. We then proceeded to determine whether, on the basis of the principles inferred from our theoretical analysis, each type of cut would benefit from the use of collocation or not.

The most obvious class in which collocation should be used is *detail*. This corresponds to the computer interface example of clicking on an object with a zoom tool to magnify or shrink it. If collocation were not found here, the cen-

tral example that has been cited as a justification for using film as informative for interface design would be disproved. A second class where collocation may be predicted is that the results of some cause or action in one shot should be collocated in the second shot (*result*): for example, an actor firing a gun in one shot and a person falling in the second. Collocation allows the two events to be linked; not collocating would require the viewer to search the screen to find the result, perhaps not seeing it before it had concluded. The third case, *following* the motion of an actor or object across the screen by cutting to different camera positions, should also generally use collocation, although there are some predictable exceptions. If the actor or object should exit on one side of the screen before the cut, another principle rules that they should be traveling in the same direction in the second shot. This requires them to be in shot following the cut or to reenter on the opposite side. Both options prevent collocation. These cuts occurred very rarely, however, and so were conservatively included within the *following* class.

Three classes involve a cut from an actor's face to something that they are looking at: a situation that can make use of eyeline cutting, in which the viewer tends to follow the direction of gaze of the actor across the screen to locate the new object of interest. Therefore, *conversation* cuts (between two or more actors who are looking at each other while they are talking) and *subjective* cuts (that show the object that the actor had been looking at) need not use collocation, but can place the object (or second actor's face) at a point between the first actor's face and the edge of the screen. We also identified a special instance of subjective cutting, which portrays the object that an actor was looking at from the point of view of a person looking over their shoulder, thus including a rear view of the actor and gives the viewer the sense of sharing their perspective. These we called over-the-shoulder cuts and felt that these should use collocation of the actor's face and the object, because of the confusion of also having the actor in shot after the cut, a potential distraction for the viewer, who might expect them to remain the topic.

Some cuts were returns to previous shots of an object or scene that had very recently been seen from exactly the same camera position (*previous*). Because the viewer would be familiar with the structure of the shot, these need not use collocation, provided that few shots had intervened (we used a criterion of no more than two intervening shots). Three classes of cut were to objects or scenes that had not recently been in shot. Where the cut was to an object whose presence was predictable from the first shot, being topically related by the narrative, collocation would aid the viewer in making the link. For a cut to a new topic, however, whose presence was not topically related to the narrative (*novel*), we felt that collocation should be avoided, lest the viewer carry over propositional information from the previous shot to the new topic. The final class of cut was that which occurred at the end of scene, with the new

shot typically being an "establishing shot" or long shot of the new scene. Again, collocation between an object on the previous shot and one in the new scene would lead the viewer to think that they might be related in some way, and so should be avoided.

## 6.1. Method

There were five participants in this experiment, all students at the University of Sheffield. A small sample is typical in this type of psychophysical study because the statistical comparisons of interest are within participants, in this case between each individual's gaze position at different times relative to cuts in a film. All participants had normal, uncorrected vision. They were paid £10 for their participation. None had seen the film before. They were told that the study was investigating eye movements made while people were watching films, but not that cuts were of particular interest. Drinking water was provided on request throughout the data collection period.

The film was presented in DVD format on a wide-screen television, viewed by participants from a chair at a distance of 150 cm, with the center of the screen approximately level with the participants' eyes. The film was presented with a widescreen aspect ratio, 58 cm wide and 24.5 cm high, thereby subtending a visual angle of approximately 22° horizontally and 9° vertically and containing the entire image as intended for the film's original cinema format, which was at a 1:2.35 aspect ratio. This visual angle is the same as would be obtained by a cinema screen 10 m wide × 4.25 m high, viewed from a distance of 25.5 m. By comparison, a 20 in. (40 cm × 30 cm) computer monitor, viewed from 60 cm, subtends approximately 38° horizontally and 29° vertically. It should be noted that the small screens of computer interfaces are small only in terms of their physical size, not in terms of the eye movements needed to search them.

Each participant was tested individually. The film was presented in its entirety (132 min, plus approximately 10 min setup time). During this time, the participant's eye movements were recorded using an Applied Science Group 4250R Eye Tracker, which uses no head restraint and tracks the participant's head position through a small magnet worn on a headband. This apparatus made the situation as naturalistic as possible for a laboratory setting. The $X$ and $Y$ coordinates of each participant's gaze location were logged in a computer file, at a rate of once per video field (i.e., every 20 msec) for the entire film. As the time of occurrence of each cut was known, eye movements following each of the 1,916 cuts in the film could be examined with this fine temporal detail. The $X$ and $Y$ coordinates were in the form of square pixels in the range 0 to 430 and 0 to 180, respectively, with each pixel subtending approximately 3 min of arc in each dimension.

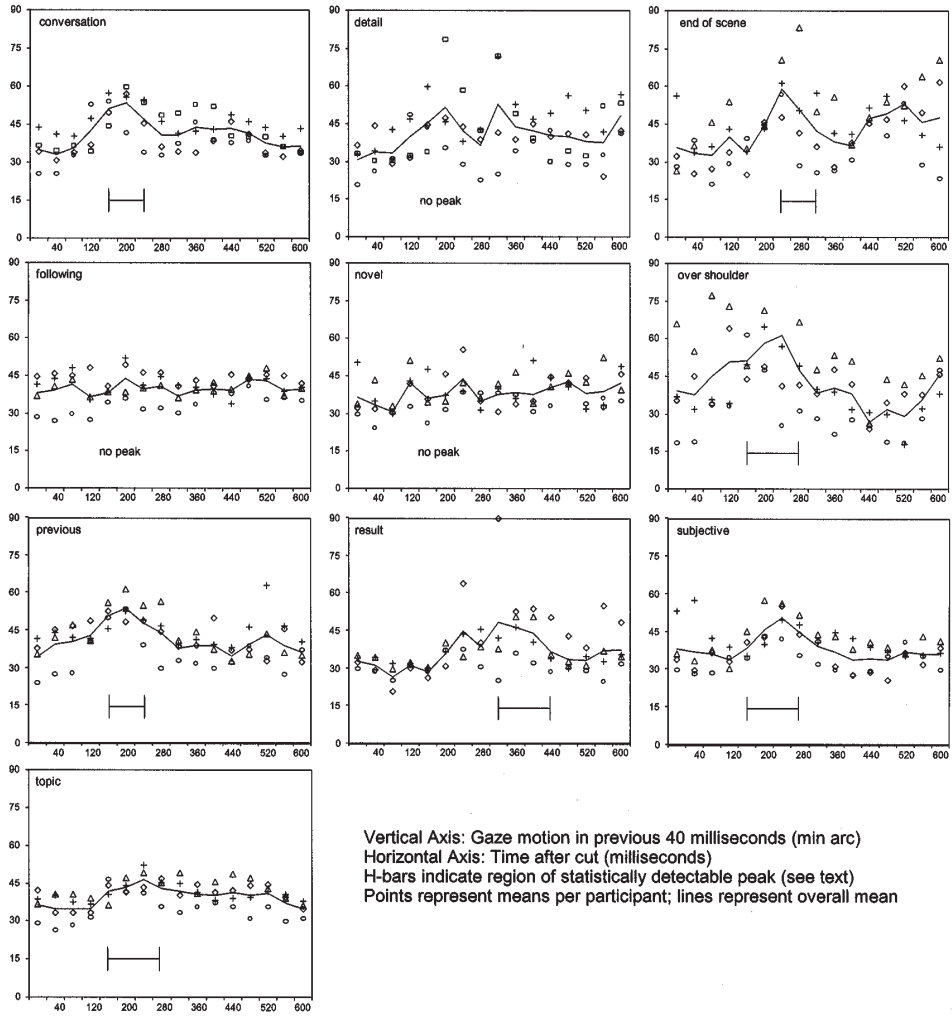
## 6.2. Results

The data for one participant were excluded from analysis, as 9.5% of the total coordinates were not in the area defined by the television screen. For the remaining four participants, any *X* and *Y* coordinates beyond an edge of the film (caused by blinks, tracking error, occasional glances away from the film, etc.) were replaced by the values corresponding to the respective edge (2.2% of coordinates being replaced, a maximum of 3.4% per participant). One hundred eighty-five cuts that were followed by another cut within 1 sec were omitted from the analysis (by proportion, mainly result, detail, or following cuts), leaving 1,731 cuts per participant.

The size of the change in gaze position in minutes of arc, disregarding direction, was calculated for 16 successive 40-msec intervals, beginning 40 msec before each cut in the film and continuing up to 600 msec after the cut. This was chosen rather than the 20 msec provided by the raw data because there are two fields of video data presented per video frame, with a new frame being shown every 40 msec (there are 24 frames per second in cine film, one of these frames being repeated in the conversion to video and DVD formats). For each cut, gaze position was in consequence measured for 16 consecutive frames, beginning with the frame before the cut and ending on the 14th frame after the cut. For each of the 10 classes of cut, mean changes in gaze location were computed for each of these 40 msec intervals for each participant. These data, and their means across participants, are plotted in Figure 9. Note that data represent eye movements since the last interval and not since the cut, and so if multiplied by 25, the vertical axis represents velocity per second. Because of the averaging process, the size of the changes are not as large as the change in gaze position for a single cut: If 40 cuts had been averaged, and each produced a 4° change in gaze position, but the changes occurred equally often after six to nine frames, the average change would be just 1° or 60 min of arc. The shape of the line is more important. Successive random changes in gaze position around a single point would result in a flat line; if gaze were completely fixed (which is not likely in eye movement data), the line would be at zero along the *x*-axis. If eye movements are being made at a systematic point in time following a cut, however, a peak in the velocity should be detectable.

One factor within participants' analyses of variance were carried out on the data for each class of cut, with the 16 measurement points used as levels of the time factor. Reliable effects were found for the following classes: conversation, $F(15, 45) = 4.712$, $p < .01$; end of scene, $F(15, 45) = 2.459$, $p < .05$; over the shoulder, $F(15, 45) = 2.712$, $p < .01$; previous, $F(15, 45) = 3.489$, $p < .01$; result, $F(15, 45) = 2.285$, $p < .05$; subjective, $F(15, 45) = 2.884$, $p < .01$; and topical, $F(15, 45) = 4.269$, $p < .01$. There were no effects of time for detail, following, and novel classes.

***Figure 9.*** **Size of change in gaze location in each 40-msec interval following the 10 classes of cut.**




Where reliable effects of time on degree of eye movement were found, the locus of the effect (i.e., when the eye movements took place) was examined with a hierarchical set of linear contrasts. This amounts to determining where statistically significant peaks are in Figure 9. The set of contrasts comprised a comparison of the first half of the data with the second half (–40 to 280 msec vs. 280 to 600 msec); then, within that first half of the data, a comparison of its

***Figure 10.*** **Percentage of each class of cut that resulted in changes in gaze location of more than 1.5° of arc within 240 msec. (White bars indicate those expected to use collocation; shaded bars indicate those not expected to use collocation.)**


first half (–40 to 120 msec) with its second half (120 to 280 msec); and so on until comparisons between adjacent intervals of 40 msec were made.

The peak amount of eye movement occurs between 120 and 200 msec after the cut for conversation and previous cuts; between 120 and 280 msec for over-the-shoulder, subjective, and topic cuts; between 200 and 280 msec for end-of-scene cuts; and between 280 and 440 msec for result cuts. None of the other contrasts examined was statistically significant, and so no peaks are identifiable for detail, following, and novel classes. These results are summarized in Figure 9, together with the *F* ratios of the relevant linear contrasts.

A constant amount of movement could represent a steady drift in gaze location across the screen or movement back and forth around a single point. To distinguish these possibilities, we calculated the distance that gaze had moved for each cut between the cut and the frame 240 msec later, the point at which most classes of cut were registering their peak amount of eye movement. We then counted the proportion of these gaze deviations that were greater than 1.5° of arc (Figure 10), this figure being taken as an approximation of the amount of movement that might be expected from saccades around a stable object.

With the exception of over-the-shoulder and novel cuts, the rank ordering of the classes supports the predictions made in Figure 8, with classes expected to benefit from collocation resulting in a lower proportion of large eye movements than other classes.

## 6.3. Discussion

The linear contrasts performed on the data identify the time intervals in which most eye movement occurred. A rank ordering of the classes of cut can be made: Eye movements follow most rapidly after conversation and previous cuts; then over-the-shoulder, subjective, and topical cuts; then end of scene; and following at the longest latency after result cuts. Because there are no statistically discernable peaks for the remaining three classes, they can be said not to induce eye movements.

This is not exactly the expected pattern of results, as listed in Figure 8. Four of the cuts producing eye movements had been expected to do so (conversation, previous, subjective, and end of scene), but over the shoulder and topical had not (however, note that these two classes were added during the classification process as special instances of subjective cuts). Although collocation (and hence low eye movements) had been expected for detail and following cuts, it had not been expected for a novel cut. It had also been expected for result, but late eye movements were found. Six of the 10 cuts behaved as we expected (accounting for 1,327 cuts in the entire film, or 69%), three did not (503 cuts), and one was ambiguous (86 cuts).

The pattern of results from the analysis of the size of gaze direction made within 240 msec of the cut (Figure 10) is broadly consistent with the eye movement data. The greatest proportion of large changes in gaze direction occurs with the previous and conversation classes of cut. Over-the-shoulder, end of scene, subjective, detail, and topical cuts are intermediate. Following, novel, and result cuts show the least proportion of large changes. The main differences are the comparatively high proportion of large changes made for detail cuts and the low proportion made for result cuts—although the point at which the latter's peak for eye movements occurred had not been reached at 240 msec. The white bars in Figure 10 indicate classes for which collocation had been expected, and it can be seen that apart from over-the-shoulder and novel cuts, the order of the classes is in line with expectations.

In summary, the different classes of film cut, as defined here, have been shown to result in different patterns of eye movement in a manner that is broadly, but not entirely, in line with the assumption that collocation would be useful in some cases but not others. Taking all of the measures together, it would appear that the classes of cut that induce the least large eye movements soon after the cut are detail, topical, following, novel, and result. Four of these were predicted to use collocation, the exception being novel cuts. Over-the-shoulder cuts were predicted to benefit from collocation, but do not seem to use it.

There are a number of factors that could contribute to differences in eye movement latency. The visual scene displayed following different classes of

cut might have different properties that affect preattentive selection of the next object. For example, the scene following an over-the-shoulder cut typically has a portion of the display occupied by an actor's back, thus making the area likely to contain the next target object both smaller and more salient, and simultaneously limiting the filmmakers' chances of using collocation. This may be why the predicted use of collocation was not detected for this class of cut. In other cases, the next target location may be highly predictable (e.g., the left–right alternation of actors' faces during extended conversation sequences). Memory may determine the target point of gaze following recognition that the shot has been seen before (previous cuts), or the target location might be cued prior to the cut by an actor's glance in a particular direction (subjective cuts). As predicted, none of these three classes of cut seem to be making use of collocation.

The latency of eye movements to a new stimulus appearing between 20 to 240 min of arc from fixation has been reported as never less than 140 msec (Ginsborg, 1953). More recently, a figure of 220 msec has been given for the execution of eye movements to unpredictable locations (Fischer & Weber, 1993). Ditchburn (1973) reported that when timing, direction, and magnitude of pulsed movements of a target are all unpredictable, latencies of eye movements are in the range of $200 \pm 50$ msec; but these latencies are reduced when the pulse is partly predictable. Latencies for the most quickly initiated eye movements (following conversation cuts) are in the 120- to 200-msec interval. We can conclude that these eye movements must use information available to the viewer prior to the cut and might even have been initiated before the cut. It may be that in choosing when to cut a shot, in these cases the editor has determined the frame by which the viewer is likely to have moved their gaze, and placed the cut there, effectively following their change in gaze rather than driving it.

This conclusion is more definite when considering that the display used in this experiment contains a wealth of complex visual information and usually several objects, rather than the simple offset and onset of dots that contributed to minimum latency figures cited in the literature. As mentioned, target locations following cuts during extended conversations are highly predictable (often comprising a series of left–right alternations), and an impending cut is invariably cued by that actor completing their utterance with the completion of sense, changes in prosody and the glances that cue turn taking so successfully in everyday conversation.

The precueing of a cut by an actor's glance in a particular direction may also be an important determinant of the latency of eye movements following subjective cuts. To examine the role of such information, the existence of such glances and their direction would need to be related to the locations of postcut selected objects and to the temporal dynamics of eye movements, a feasible

enterprise that is beyond the scope of this study. At a more general level, the timing of cuts (rather than the locations of objects) may be cued by preceding shots being in some way complete (e.g., a fight scene finally having a victor or a moving car coming to a standstill). Cuts are not directly cued in such circumstances but do follow with high probability as this is part of the conventional language of narrative filmmaking.

Predictions about collocation with topically related and novel topics were clearly in the wrong direction. It had been expected that collocation should be used when the objects of interest before and after the cut were topically related, but that it should not be used when the novel topic could not be predicted from the narrative. The data indicate that the reverse is happening: Novel cuts induce one of the lowest amounts of eye movement; topical cuts induce more. Here the analogy between film and computer interface tasks may be helpful in the reverse direction: Unexpected windows and alerts are commonly designed so that they are unavoidable, opening "in front" of the user's current focus. Expected task-related dialogs often appear to one side so that the focus to which they refer is not obscured. It may be that when the film viewer can anticipate the object of interest after the cut, they can also anticipate a noncollocated position; certainly, the predictable identity of the related topic makes the subsequent visual search easy. The relevance of a novel topic, however, is less apparent, and so using collocation removes the need for the dual task of visual search plus assessment for narrative relevance. After all, if an unexpected object appears in an unattended position, it is unlikely to be noticed; if it is necessary for it to be noticed, it had better be placed where the viewer will see it. This would be more important than avoiding the carryover of propositional information suggested by our theoretical model.

The strongest claims about the benefit of collocation were made for detail, result, and following cuts; and all of these three classes do seem to have resulted in the fewest fast eye movements, although the data for detail cuts is the most noisy. Taken at face value, this confirms May and Barnard's (1995) suggestion that computer interface designers should take into account the relative positions of the expected locus of visual attention before and after a window has been opened. When cutting in or out from a view, the detail that has been operated on (or selected) should be collocated. When an object has been operated on, any view of the results of this operation should be collocated with the focus of the operation, even if they are displayed in a new window. When an object that is selected or being operated on is moving so that it will soon move beyond the window, the point of view represented should be changed (corresponding to a change in camera position) to avoid this, keeping the object collocated. This last possibility is the hardest to exemplify in applications that use two-dimensional scenes, such as word processors or graphical editors; representations of three-dimensional scenes, such as computer-aided design applications, data

navigation, and even virtual reality devices might profit from the application of this common filmmaking technique.

## 7. APPLICATION TO INTERFACE DESIGN

In Ridley Scott's (1982) film *Blade Runner,* the title character is searching for information in a high-resolution hologram snapshot. He selects a point within the picture and instructs the computer to expand the resolution—but instead of a smooth zoom in or a cut straight to the highest resolution, the computer makes a series of four or five cuts, each one of progressively greater resolution. Because this is a completely fictional situation, there are no technological constraints on the design of this interaction. Although it could be argued that the director wants to impress on the viewers this sequence of processing that is being carried out on the image, the discrete steps into the detail of the image also make it readily apparent just what is going on, both for the narrative of the film (the character presumably knows what his command meant, but the viewers might not) and for the character as a user of the system—although he might know what command he gave, he still has to be able to comprehend the results.

Ridley Scott's experience as a film editor could have prompted him to use a single collocated close-up, but because the viewpoint of the image also had to rotate slightly to reveal a previously occluded element in the snapshot, he realized that this would not work. Instead he chose a series of close-ups, each of which shifted the viewpoint by a small amount, keeping the distortion of the predicate structure in the view to manageable proportions. Although the interrogation of hologram snapshots is still future technology, three-dimensional animations are currently used for the display of complex data structures, and this issue of motion through the representation is critical.

The recommendations that we have made for the construction of film cuts, involving the collocation and translation of the psychological subject and the maintenance or otherwise of the predicate structure of the scene, can be applied to the design of computer displays. It may seem obvious that, on changing a display from one representation to a close-up of part of that representation, the element that the user is working on should be kept within the screen; but as described in the introduction, not all current graphical editing packages follow even this simple rule. Some make the center of the current image the center of the new image, so that elements that were around the edge of the display vanish and the user must scroll to bring them back into view—and if the center of the image was blank, the resulting blank screen is not much use in helping the user to decide the direction to scroll. In multiwindowing environments, it seems that no consideration at all is given to the adverse consequences of collocation on the equivalent of end-of-scene cuts—the closing of windows or their replace-

ment on switching from one application to another. The result is that one window's contents is replaced on the screen by what was previously behind it, regardless of the lack of propositional and implicational correspondence.

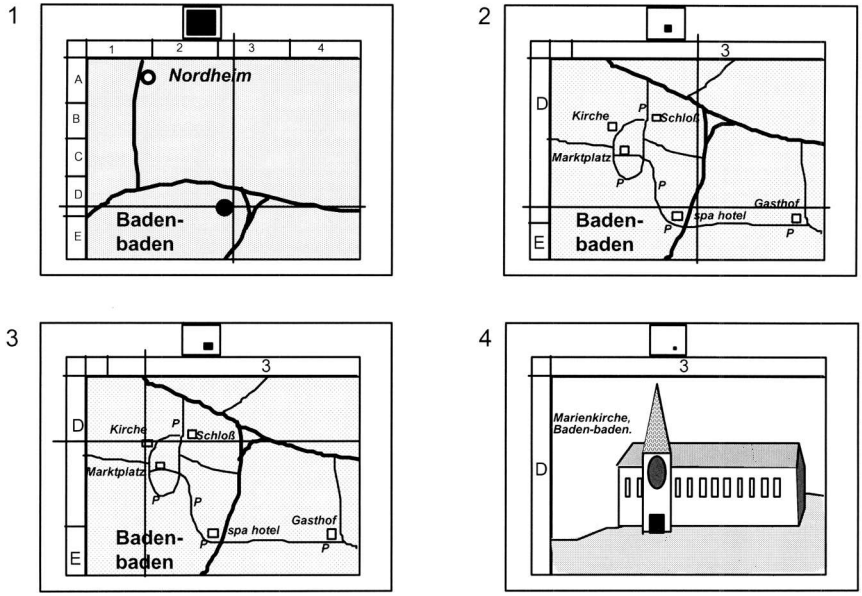
It is a common assumption that in moving through a three-dimensional representation, it is an advantage to use a continuous transformation of the display so that objects being approached gradually increase in size and resolution, with objects that are being passed by smoothly slipping out to one side or other of the display, such as the "walking metaphor" and "logarithmic flight" of Robertson, Card, and Mackinlay (1993). Such interfaces certainly create a vivid illusion of motion in the viewer, sometimes to the point of inducing motion sickness or unpleasant after effects. Where the user just wants to get more detail about a particular element, our model suggests that a simple cut to a close-up view would be sufficient; perhaps as in Ridley Scott's (1982) example, over in a few discrete steps to recreate the effect of a transition in the object representation to the element's visual substructure.

Similarly, in moving from close-up displays of one region to another, our analysis of visual transitions through a structural representation of the scene suggests that it may be best to do so via a smaller scale overview that encompasses both regions. Collocation could be used on the opening-up cut, with a pan across the view to bring the new destination to the center (or some other cue to make it the subject of the user's object representation), followed by a collocated close-up to the relevant part of the new region. This would allow both the object and propositional representations to be formed without difficulty. An example is the "small screen map" discussed by Barnard, May, and Green (1991). In this scenario an on-screen information map was being developed for the general public to find their way around, to get an overview of the general area, and to identify key detailed points. These tasks require different levels of detail; due to the small screen being used, the whole map could not be shown at the highest required magnification.

Three design alternatives were suggested, but these all tried to avoid switching between displays on the assumption that users would benefit from having both scales of map available simultaneously. It should be clear from our model of the cognition involved in the interpretation of a visual scene that people cannot focus on information from disparate parts of the visual scene simultaneously.

A design that was suggested following this realization replaced the visual transitions that the user would need to make between different areas of a single display with changes in the screen display (Figure 11). Instead of the user having to attend to a different part of the display, the whole display would be taken up by the new scale view. To move to another detailed view, the user would first move back out to the overview and indicate the new point that they wished to focus on.

***Figure 11.*** **The small screen map interface, showing four successive screens. The user clicks on Baden-baden (1) and there is a cut to a close-up showing a street map of Baden-baden (2)—the new display uses the shape of the major roads to preserve the immediate predicate structure around the center of the crosshairs and adds in the minor roads. The user then moves the crosshairs to the Kirche (3) and clicks again for a further close-up (4)—the new display uses collocation to position the name of the view appropriately (the location would depend, of course, on the position of the crosshairs in the preceding display controlled by the user).**




One of the original design suggestions was to present the overview on two thirds of the screen and have a detailed view in the remaining one third. Although these areas would remain in fixed positions on the display, the user could move the logical position of the detailed view (and hence its contents) around the overview, where it would be represented by drawing a box or reversing the colors of the area represented. In addition to the problems of making a visual transition between the elements in the overview and the detailed view that we have described, this suggestion also faces the same difficulties as the portrayal of a conversation in the front seat in a car (Figure 7)—as the detailed view was moved in one direction across the overview, elements within it would smoothly scroll in the opposite direction. If the user tried to look from one view to the other, the contrasting flow of elements in the predicate structures would make it very difficult to relate the two representations.

These examples have dealt with situations where the user is driving the interaction, and the display is responding to their commands. This is apparently quite different to the narrative style of films where the user is passively led through the sequence of shots, and yet in our examples the visual composition of the display has been actively determined by the computer to facilitate the user's comprehension of the scene—the degree of change between the successive cuts into the holographic snapshot and the positioning of the text labels in the small screen map scenario, for example. Interactions where the computer might take an even more active role, such as in process monitoring displays, could also benefit from the application of these techniques to alert the user to important events. They could also provide a basis for the design of dynamic displays where essentially static displays are currently used, for instance in videoconferencing—one participant watching two others converse would benefit from having an "eyeline match" created for them, rather than having them both talk directly to the camera.

Experimental findings with the hydra prototype videoconferencing system support this conclusion. The hydra system provides a single 8 cm screen for each participant, mounted above a camera in a desktop display unit. A number of these desktop units can be arranged in each conference location to represent the distant participants. As participants turn their head to look from one unit to another, their images in the remote locations appear to look toward the appropriate participant's unit. Sellen (1992) reported that people preferred this system to a single screen with multiple windows because they could tell when other people were attending to them and could selectively attend to other individuals.

Another important difference between film and interface displays that has been mentioned in passing is that the scene represented on the film screen is usually of a single, coherent physical setting, whereas an interface display often represents views of several unconnected applications or processes. Film viewers seem quite able to watch a scene of, say, a city street without difficulty despite many different events happening simultaneously (people and vehicles moving in several different directions and lights flashing on or off), but an interface with these properties would seem very busy and distracting. Our analysis of the visual structure allows us not only to explain why this is but allows us to make recommendations about making the interface display's "busyness" manageable. The propositional representation of the filmed street scene may have a lot going on, but the elements and their attributes are all consistent with a single schema, and can hence be contained within a single implicational representation. In film terms, they are part of the same narrative. If the viewer does make one of the background elements of the scene the subject of their propositional representation (due to its visual salience, perhaps), they can do so without requiring a lot of PROP→IMPLIC and IMPLIC→PROP transformations to derive a new

implicational representation to comprehend it. They can also make a transition back to the appropriate element of the scene without difficulty. This is clearly not the case with an interface display where a spreadsheet may be partially hidden by a word-processing window, with a video image running in another corner and an "incoming mail" flag flashing on the menu bar.

An example of a problematic multiwindow display arose at a workshop involving commercial interface designers, where we were asked to model some interfaces that had been brought along by the designers. One of these was a view of a screen from an electron microscopy analysis package with several different windows. One window contained a view of a microscopy sample, others the results of an analysis of the constituents detected in the sample, but represented in different ways (as pie charts, histograms, tables, etc.). Each window had its own menu bar, although these were all identical; there were several display elements that were common to each window (e.g., the atomic symbols). As the pointer was moved across the view of the sample, a histogram of the analysis would update in another window. Although many of the display elements thus had a "common fate" and their representations were interdependent, their separation into different windows made this difficult to determine. While trying to understand the interface design, in fact, we worked through the substructure of each window, verbalizing the propositional representations that we could form, until on about the third window we were able to form the implicational recognition that they were all really representations of the same information—whereupon our verbalization became, "Oh, right!" The problem of making the link between a particular graphical or tabular portrayal of the data and the histogram that it had been derived from could be simplified considerably by showing them in the same window—unifying their object representations, and so making explicit the proposition that they were based on the same sample.

To be fair, this screen display was a little more cluttered than it would be in practice, because it was an "advertising shot" taken to show all of the possible functions simultaneously. In practice the users would not call for a pie chart and two different tables for the same data simultaneously—but as an advert it was probably counterproductive in that it gave the impression of a rather confusing interface. The desire to show off all of the attractive features of an interface at once is often apparent in displays of multimedia technology, particularly where the machine is able to show more than one moving image at a time. Although this feature does have useful application in multipoint videophony, for example, or in allowing two colleagues to collaboratively watch or edit a video clip, the displays seldom reflect the implicational unity of these applications. Instead, one image will zoom in while another will zoom out and a third is panning across a different scene entirely. The experience of trying to look at these displays is not pleasant.

## 8. UNDERSTANDING SCENIC AND STRUCTURAL CHANGE

To make a start at understanding how the attributes and location of the psychological subject of a scene (and its predicate structure) can be manipulated in computer displays, we can identify three broad types of dynamic change in the scene:

1. If the display changes (e.g., a new window replaces the previous focus) and offers a new structure that has an element located close to the preceding subject, this will become the user's new focus of processing. This is a transition by collocation.
2. If the display changes and the new structure does not contain an element that is located close to the preceding subject, the user will establish a new focus either by retopicalizing on a translation of the previous subject (i.e., in a new location on the screen) or on another significant element. Any new subject will be determined by the salience of the elements of the new structure and their proximity to the previous focus.
3. If the display does not change entirely but the structure is altered by the repositioning of elements, or the alteration in the attributes of some elements (e.g., brightness, size, or color), then the user may make an involuntary transition to a new focus.

In these three generalizations we distinguish between a *scenic change,* where the complete structure changes (1 and 2); and a *structural change,* where elements of the structure move or their attributes are altered (3). This last type corresponds to the jump cut, which is generally regarded as unfilmic because of its propositional consequences for the narrative, but which for the same reasons may be valuable in a computer interface because it serves to interrupt processing and attract the user's attention to the incongruous element of the display. Just like the balls in Figure 3, the degree of change in the element's attributes will determine whether they appear to be consistent objects whose appearances vary or whether they appear to be appearing and disappearing, and hence more likely to force a retopicalization of the user's object representation. These questions must remain open for now, but could potentially be resolved empirically.

In conclusion, it is possible to describe the effects that have been developed to direct and maintain the film viewer's comprehension by adopting a reasonably concise set of rules concerning the construction of the screen image. These rules could be tested with a methodology that examines the ease with which people can comprehend simple sequences of images that either accord with or

contradict them. These scenes would have to be constructed explicitly to violate filmic principles; so rather than obtaining them from existing commercial films, as we have done here, the material would have to be produced explicitly for the investigation, with the participation of professional filmmakers, and preferably embedded within a reasonably realistic narrative. Computer displays that change dynamically are no different in principle because users must comprehend the changes that have taken place, and so should also benefit from the understanding of display construction that these rules can provide. Here the empirical work is more amenable to laboratory investigation within a conventional HCI environment, perhaps using eye tracking equipment to test our predictions about the timing of changes in gaze location and the occurrence of display changes. Further work could then be directed toward the mapping of other filmic principles to interface design, and perhaps even to work on the modeling of narrative construction.

## NOTES

***Support.*** This research was supported by the European Commission Training & Mobility of Researchers network "TACIT" and by UK EPSRC Grant GR/M89331/01.

***Authors' Present Addresses.*** Jon May, Department of Psychology, University of Sheffield, Western Bank, Sheffield UK. E-mail: jon.may@shef.ac.uk. Michael Dean, King's College Hospital, Denmark Hill, London, UK. E-mail: michael.dean@kingsch.nhs.uk. Philip J. Barnard, Medical Research Council, Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge UK. E-mail: Philip.Barnard@mrc-cbu.cam.ac.uk.

***HCI Editorial Record.*** First manuscript received May 16, 2002. Revision received May 20, 2003. Accepted by Stephen Payne. Final manuscript received June 23, 2003. — *Editor*

## REFERENCES

Allen, W. (Director). (1986). *Hannah and her sisters* [Motion picture]. United States: Metro Goldwyn Mayer.

Balázs, B. (1970). *Theory of film (1945).* New York: Dover.

Barnard, P. J., & May, J. (1999). Representing cognitive activity in complex tasks. *Human–Computer Interaction, 14,* 93–158.

Barnard, P. J., May, J., & Green, A. J. K. (1991). *Preliminaries for the application of approximate cognitive modeling to design scenarios* (Esprit BRA3066 Amodeus working paper RP4/WP11). Cambridge, UK: Medical Research Council–Applied Psychology Unit.

Bernstein, S. (1988). *The technique of film production.* London: Focal Press.

Boorstin, J. (1991). *The Hollywood eye: What makes movies work.* New York: HarperCollins.

Campbell, M. (Director). (1998). *The mask of Zorro* [Motion picture]. United States: Columbia Tristar

Carroll, J. M. (1980). *Toward a structural psychology of cinema*. The Hague, Netherlands: Mouton.

Carroll, J. M., & Bever, T. G. (1976). Segmentation in cinema perception. *Science, 191,* 1053–1055.

Ditchburn, R. W. (1973). *Eye movements and visual perception.* Oxford: Clarendon.

Eisenstein, S. M. (1972). *Film form* (J. Leyda, Trans.). New York: Harcourt Brace. (Original work published 1949)

Fischer, B., & Weber, H. (1993). Express saccades and visual attention. *Behavior and Brain Sciences, 16,* 553–610.

Frith, U., & Robson, J. E. (1975). Perceiving the language of films. *Perception, 4,* 97–103.

Ginsborg, B. L. (1953). Small involuntary movements of the eye. *British Journal of Ophthalmology, 37,* 746–754.

Glenney, M., & Taylor, R. (1991). *S. M. Eisenstein: Selected works, Volume 2: Towards a theory of montage.* London: British Film Institute.

Hochberg, J. (1986). Representation of motion and space in video and cinematic displays. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (Vol. 1). New York: Wiley.

Katz, S. D. (1991). *Film directing shot by shot.* Studio City, CA: Michael Wiese Productions.

Kraft, R. N. (1986). The role of cutting in the evaluation and retention of film. *Journal of Experimental Psychology: Learning Memory and Cognition, 12,* 155–163.

Kraft, R. N. (1987). Rules and strategies of visual narratives. *Perceptual and Motor Skills, 64,* 3–14.

Kubrick, S. (Director). (1968). *2001: A space odyssey* [Motion picture]. United States, Warner Brothers

Lindgren, E. (1963). *The art of the film* (2nd ed.). London: Allen & Unwin.

Maltby, R. (1995). *Hollywood cinema: An introduction.* Oxford, England: Blackwell.

Mamet, D. (1991). *On directing film.* London: Faber & Faber.

May, J., & Barnard, P. J. (1995). Cinematography and interface design. In K. Nordby, P. Helmersen, D. J. Gilmore, & S. Arnesen (Eds.), *Human–computer interaction: Interact 95* (pp. 26–31). London: Chapman & Hall.

May, J., & Barnard, P. J. (2003). Cognitive task analysis in interacting cognitive subsystems. In D. Diaper & N. Stanton (Eds.), *A handbook of task analysis for HCI* (pp. 291–325). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Metz, C. (1974). *Film language: A semiotics of the cinema* (M. Taylor, trans.). New York: Oxford University Press. (Original work published 1971)

Münsterberg, H. (1970). *The film: A psychological study (1916).* New York: Dover.

Richards, R. (1992). *A director's method for film and television.* Boston: Focal Press.

Robertson, G. G., Card, S. K., & Mackinlay, J. D. (1993). Information visualization using 3D interactive animation. *Communications of the ACM, 36*(4), 57–71.

Scott, R. (Director). (1982). *Blade runner* [Motion picture]. United States: Warner Brothers

Sellen, A. J. (1992, May). Speech patterns in video mediated conversations. *Proceedings of SIGCHI Conference on Human Factors in Computing Systems,* 49–54. New York: ACM.

Taylor, R. (1988). *S. M. Eisenstein: Selected works, Volume 1: Writings 1922–34.* London: British Film Institute.

Teasdale, J., & Barnard, P. J. (1993). *Affect, cognition & change.* Hove, UK: Lawrence Erlbaum Associates, Inc.

Young, E., & Clanton, C. (1993). *Film craft in user interface design.* Tutorial presented at the *InterCHI 93* conference, Amsterdam, The Netherlands.