Psychocinematics

EXPLORING COGNITION AT THE MOVIES

Edited by Arthur P. Shimamura
COGNITIVISM, PSYCHOLOGY, AND NEUROSCIENCE:

MOVIES AS ATTENTIONAL ENGINES

Noel Carroll & William P. Seeley

The Power of Movies

The folk intuition about the motion picture is that the power of the medium emerges from the strong, putatively realistic grip of ordinary movie-going experience—an experience that allegedly amounts to an illusion of reality. This intuition is reflected in the pervasive emphasis on realism in film theory. However, movie-going experiences for most of the movies that most people see are scarcely realistic. Moreover, it is to a large extent the ways in which movie-going experience departs from the subjective phenomenology of ordinary experience that, in fact, account for its power (Carroll, 1985). We argue, therefore, that the realist intuition in film theory is an odd intuition, despite its perhaps surface plausibility. We start with a distinction. The term *movies*, as we will use it, is a narrow term that refers to the kind of mass media narrative motion pictures associated with big Hollywood studios, Bollywood, and indie distribution houses that feed mainstream “art houses” like the Angelika Film Centers in New York, Houston, and Dallas.1 *Motion pictures*, on the other hand, is a broader term that refers to the medium as a whole and includes movies, art films, experimental films, flip books, hand-etched animations made frame by frame with an Exacto knife from a loop of old celluloid, and whatever else under the sun that can be fit into the category. In this sense the term *movies* denotes a motion picture genre. The most striking feature of this genre is its cross-cultural capacity to gather viewers up into a narrative and thereby deliver intense engaged experiences. We can call these two aspects of movies their *widespread accessibility* and *widespread intensity*, respectively. It is certainly true that any theory of motion pictures worth its salt owes an explanation of the qualitative, experiential grip of movies these terms describe. However, film realism is the wrong place to look for such an explanation.

---

1 This category is not limited to movies but also includes network television productions.
The widespread intensity of our experiences of movies and the widespread accessibility of their content are likely sources for the intuition that movies are experienced as illusions of reality. Our relationship with and to the characters and events on the screen can feel strikingly compelling. It may be that the only way some viewers believe they can explain this reaction is to invoke the notion that we experience movies as we experience life outside the cinema. That is, we must be undergoing an illusion of reality. However, the kinds of approach and avoidance behaviors common to our awareness of the presence of danger in the ordinary environment are conspicuous in movies in our affective and emotional responses to movies. We are not literally paralyzed by fear at the movies when the Fifty Foot Woman ravages her neighborhood. We are delighted by it, even as it raises our anxiety levels. Of course, some of us may cry an awful lot at the movies, but we don't attempt to alleviate the suffering of characters in the way we might in ordinary experience. That is, a moviegoer's behavioral responses at the movies, even when fully immersed in emotion, are radically divergent from what one would predict if he or she were under the thrall of the illusion of reality. Moreover, most of the world passes us by in ordinary contexts without our ever taking notice. Not so at the movies. Here every detail announces its salience like a klaxon. We take heed of everything that is significant in the movie narrative. We should be so lucky as to be so perspicacious in everyday affairs. In short, a great many of our recurring responses to movies are so out of sync with the behavior we exhibit in ordinary contexts that it is unlikely that they can be explained in terms of the experience of a so-called "illusion of reality." The explanation must lie elsewhere. However, before we introduce the account we favor, let us take a closer look at some competing theories.

THE ILLUSION THESIS

Film realism can loosely be defined as the claim that the conventions of the theater conspire with the compositional structure of movies to induce an illusion of reality in our movie-going experience. The ordinary environmental distractions that form the backdrop of everyday behavior are masked in the theater. The lights go down, conversations subside, the soundtrack comes up, and as we become enveloped by the darkness we are literally drawn into the events that unfold on the screen before us. We can call this the illusion thesis. The trouble with this view is that it is hard to make sense of the meaning of the term illusion in a way that preserves realist intuitions (see Carroll, 2008).

Illusions come in two flavors: cognitive and perceptual. Cognitive illusions are cases in which we believe that what we see is true—we are in the cognitive grip of the illusion. We are in the throes of these kinds of illusions to the degree that we don't notice their inconsistency with the natural environment around them (e.g., scale illusions involving far-away objects). The trouble is that once we do notice we can never regain the sense of epistemic innocence that makes these illusions so compelling. This is a problem for putative intuitions about our engaged experience at the movie house. There are just too many obvious cues that count against this kind of epistemic naïveté to genuinely count movies as cognitive illusions. We don't really seem to have the experience of really seeing events unfolding in front of us—we don't treat the depicted events as a slice of our present reality. We don't take out our cell phones and call the police when we witness a murder in a movie. We are quite aware that we are sitting in a (likely slightly shabby) theater, surrounded by strangers, looking at a flat reflective screen. We may even consciously look out for the anachronistic cue marks signaling a change of reel.

What about the notion that movies constitute perceptual illusions? We are familiar with many examples of perceptual illusions. A straight stick appears bent when it is stuck into water. One of the lines in the Mülller-Lyer illusion appears longer than the other. Pavement in the distance often appears wet on a hot summer day. And so on. In these cases it doesn't matter that you recognize that what you see is an illusion. The experience of the illusion is unchanged by your belief; the processes are, as philosophers and cognitive scientists say, cognitively impenetrable. The result is that we experience realistic psychological effects in the presence of these kinds of stimuli and environments—effects so stable and strong that no matter how hard we try we just can't shake them.

Suppose, then, for the sake of argument, that there are some aspects of movies that constitute perceptual illusions. What aspects might they be? At best, it may be alleged that we really do see movement, depth, objects, and dynamic events depicted on the screen (as opposed to shifting two-dimensional patterns of light or a series of still shots projected in rapid sequence). But, even if this were uncontroverted, the concession would be a small victory. The realist hypothesis isn't that movies are primarily illusions of movement, space, or depiction. It concerns our qualitative engagement with the depicted events, actions, and characters. It is the claim that a putative belief in the drama before us is what keeps us glued to the screen. The pertinent question is, therefore, "Are we really under the perceptual illusion that we are actively experiencing the events we perceive unfolding on the screen?" The answer seems to be no. The relevant experiences don't have the right structure to support that interpretation. The problem lies in our perspective on the events unfolding in front of us. Ordinary perceptual experience is marked by an egocentric perspective. As we move around in the world our perspective on events and objects changes in a way that reveals our position and orientation relative to them. This information is conspicuously missing from the movies. So, even if in some sense our eyes are fooled, our bodies aren't. And even our eyes really aren't fooled. Accommodation, a phenomenon wherein muscles in our eyes adjust the shape of the lens to maintain focus as we shift attention back and forth between objects perceived in depth, is tuned to the flatness of the screen, not the depth of field of the image. It has to be in order for us to maintain a clear image of the depicted scene. And, of course, as discussed earlier, whatever embodied affective, behavioral responses mark our intense engagement with movies, they are the wrong ones. We don't flee danger at the movies, we don't attempt to help the injured, nor do we try to console the bereaved. No matter how you slice it, explanations of the qualitative, experiential grip of movies that appeal to an illusion of reality are dead in the water.

THE FILM LANGUAGE HYPOTHESIS

The failure of folk intuitions about film realism leaves us without an easy explanation of the widespread accessibility and intensity of movies. One way to address this problem is to look
under the hood to see how movies work as communicative devices, to evaluate the nature of a consumer’s engagement with the genre. Semiotic theories have provided an influential approach to this question within film theory. On this account shots are treated like words, cinematic sequences are treated like sentences, and editing is treated as a stereotyped language of stylistic devices. The job of the viewer is to read sequences of shots or to recover the content of a movie from its putative surface grammar. Movie making can thereby be interpreted as a particularly gripping type of storytelling whose widespread accessibility is explained as an artifact of its quasi-linguistic structure. We can call this the film language hypothesis.

There is, as with film realism, a surface plausibility to the film language hypothesis. Sentences are built up systematically out of meaningful grammatical units (e.g., nouns and verbs). The meaning of a sentence is a product of how these units are combined. Stylistic variance in the way words are used and sentences are constructed is a familiar way to articulate the meaning of sentences in poetry, literature, and everyday speech. Likewise, cinematic sequences are composed of chains of shots that yield more complex meanings when combined and juxtaposed. Stereotyped camera and editing techniques are the primary means a movie maker uses to articulate the content of shots and sequences. Further, knowledge of the stereotypical film editing styles of different directors contributes to our capacity to read the way shots and scenes have been articulated. Therefore, it has been argued that cinematic sequences are built from sentence-like codes governed by syntax-like production rules that are used to articulate their content.

However, there are critical dissimilarities between film and language that challenge the validity of the film language hypothesis (Currie, 1991). For starters, shots aren’t like words. Words are abstract symbols. The relationship between a word and its referent is, as a result, arbitrary—there is no natural relationship between the form of a word and the form of its referent. This entails that the meaning of a word needs to be learned. Motion picture shots are different. They are pictorial representations. Dog pictures depict dogs because they resemble dogs in salient—albeit rather abstract—ways that are sufficient to prompt the recognition of dogs in normal perceivers. We can recognize a camp fire or a forest fire in a photograph of smoke rising from a strand of trees, and discriminate between photographs of the two, because smoke is an effect of fire and so reliably indicates its presence. More importantly, all that is required to recover the content of a picture or a shot is the natural perceptual recognition capacities that we all share by virtue of ordinary developmental processes. Certainly some perceptual learning is required—we need to be able to categorize a stimulus as belonging to a familiar class of objects in order to recognize it. But the learning involved is nothing like language learning. If I can recognize a dog, I can recognize a dog in a picture that depicts one. In fact, if I can recognize a particular kind of dog, I can likely recognize a picture of any animal in that broader category as a dog picture. There is nothing over and above learning to see one that I need in order to perceptually recognize the other. Language is nothing like this at all.

There is no natural relationship between the word samoyad and samoyads. The capacity to recognize samoyads doesn’t naturally translate into an understanding of samoyad, nor does an understanding of the meaning of the word samoyad naturally confer a capacity to recognize samoyads (although a sufficiently articulated description of the associated perceptual category would help). Likewise, the words samoyad and dog bear no natural relation to one another. Therefore, the capacity to understand the first does not necessarily generalize to other members of the latter category. All of these particular semantic relationships have to be learned.

The analogy between cinematic sequences and sentences fares no better. The meaningfulness of a sentence emerges from the way syntactic rules are systematically used to combine smaller grammatical units—nouns, verbs, and phrases—into meaningful wholes. Violations of these rules produce nonsense sentences. Cinematic sequences are, likewise, constructed from spatiotemporally discontinuous shots. However, the editing rules used to stitch these smaller units together are not likewise hard and fast rules like syntactic rules. They are, rather, artistic rules of thumb that have been developed to help maintain and enhance the unity of, the perceived coherence and continuity within, these sequences. Furthermore, sequences that violate these conventions can be just as meaningful as those that adhere to them. In fact, violating artistic rules of thumb for effect is itself an established means to articulate content, a compositional strategy that draws attention to itself in a way that prompts a viewer to consciously reflect on what the movie maker is up to (i.e., on what it might mean to have composed a shot that way in that context). Examples of these kinds of violations in movies are well known. The 360-degree rule is a convention that states that the camera should be kept on one side of the action so as not to mix up the direction of movement. John Ford violates the 360-degree rule in his 1939 movie Stagecoach in order to enhance the tension in a critical chase scene. Likewise, Jean-Luc Godard employs a 360-degree pan in Weekend to enhance the sense of disorientation in his depiction of decadent middle-class French culture. Godard also consistently violates the 30-degree rule—the convention to shift the camera angle at least 30 degrees between shots to avoid jump cuts—in Breathless in order to reinforce the improvisatory, "spontaneous" feel of the movie. If editing rules were like syntactic rules, scenes like these would be incomprehensible. But they aren’t. Why not? Understanding their contents just isn’t a matter of quasi-linguistic film competence. Rather, we perceive the depictive content of cinematic shots and sequences in large measure just as we do everyday objects and events in our ordinary environment.

Of course, a familiarity with film conventions may be needed to understand and evaluate the reasons particular editing patterns are used the way they are in a movie, in order to sew what is seen together into a coherent narrative. However, if the language of movies is indeed a language, then we must suppose that it is surely a second language. Learning a second language is a labor-intensive task. It takes an enormous amount of training and practice. We don’t say that we have truly learned a second language until we become fully immersed in it—until we learn to stop translating and genuinely think in that language on the fly. Not so for movies. We immediately comprehend the depictive content of shots and sequences using nothing but our natural perceptual recognition capacities. This is an important point to keep in mind. The trouble with the film language hypothesis is not just that there is a dissimilarity between language and film. The target of this discussion is the universally intense cross-cultural experiential grip of movies, their widespread accessibility and intensity. If film comprehension depended on a quasi-linguistic communicative code that had to be learned like a second language, we would be hard pressed to explain these phenomena.3

3 Perhaps more critically, linguistic processing and perceptual processing draw on different networks of brain areas. Other than studies focused on the comprehension of movie dialogue, we know of no studies that link the comprehension of other shots or edited sequences with dedicated linguistic processing areas. Although we know of no studies demonstrating that there is not any such dependence, we suggest that the range of dissimilarities between movies and language described in this section render it unlikely. Rather, our claim is that the film language hypothesis is best treated as a metaphor, and a poor one at that.
Cognitivism

Cognitivism emerges as an alternative that lies both between and outside film realism and the film language hypothesis. Neither of the latter alternatives adequately captures the basic structure of our engagement with mainstream movies. Yet there is something right in each. On the one hand, the qualitative experiential grip of movies is grounded in the way we perceptually engage with them, the way we come to recognize, understand, and appreciate their content. On the other hand, the tools of the medium are stereotyped stylistic rules of thumb that are used to deliver information sufficient for an audience to recognize, recover, and reconstrue the depicitive, narrative, and artistic content of a movie from the surface structure of its shots and sequences. Cognitivism is grounded in the claim that movies are pictorial representations that carry information sufficient for viewers to access their visual content using nothing more than their natural perceptual recognition capacities. However, movies differ in critical ways from our ordinary perceptual environment. They are framed by and encapsulated within a depicted space discontinuous with our own. They do not carry egocentric information about the relative orientation of a viewer to their depicitive content. And, more importantly, movie makers are constantly changing and revising our point of view on the action in ways that defy the construction of a unified spatiotemporal point of view. Cinematic sequences, rather, present a perspective that is spatiotemporally discontinuous and perceptually sparse.

Counterintuitively, it is the practice of continuously revising the point of view of the viewer that accounts for the widespread intensity, the experiential grip, of movies. Camera movement, lens movement (e.g., zooms-in), and editing techniques are stereotyped formal devices movie makers use to control attention and filter the information available to the audience over the course of a movie narrative. In this regard, movies can be thought of as attentional management devices, tools intentionally designed to direct a viewer’s attention to their artistically salient perceptual features, perceptible features responsible for their affective, depicitive, narrative, and semantic content. The net consequence is that movies hold the attention of viewers in their grip, focusing, shaping, and carrying their perceptual experience through the twists and turns of the story line. This, in turn, is a means to mold the narrative in the mind’s eye of a viewer and direct the viewer’s emotional engagement with characters and events.

We argue that movies are attentional engines fine tuned to a range of natural cognitive and perceptual capacities. The key to the initial development of our model is the claim that the basic building blocks of film, the shots from which scenes and narrative sequences are constructed, are recognitional prompts. A recognitional prompt is a stimulus sufficiently structured to enable a perceiver to identify its depicitive content by matching its formal/compositional content to declarative knowledge of the structure and function of target objects and event types. Recognitional prompts of this sort can be pictorially replete or pictorially sparse (see Goodman, 1969). For instance, color digital journalistic photographs of the sort found on the front pages of major newspapers carry a phenomenal amount of detail about their subjects. In contrast, a simple gesture, a thin line curved just right, may more than suffice in caricature. This kind of pictorial variance is artistically productive. When a visual artist or movie maker chooses how to construct a pictorial representation, he or she can, indeed, must, choose just how replete or sparse a recognitional prompt the image will be. This choice is constrained by the capacities of the medium and the communicative goals of the artist, by decisions about how to best articulate the intended content of the work.

The pragmatics of image structure in pictorial representations mimic the pragmatics of ordinary perception. Given that cognition is a limited and time-consuming resource, our capacity to smoothly and efficiently orient ourselves to a dynamic environment in goal-directed contexts requires selectivity. Sometimes all we want to do is register that something is there in the environment. At other times we want to know in detail how certain aspects of the environment are oriented to ourselves and how their parts are structured. But we rarely, if ever, need to generate a global spatial model of the local environment. A capacity to locate and identify target objects and engage those aspects of the environment that we need to interact with in order to accomplish our goals will suffice; for example, if I want to shake my brother’s hand I do not need an explicit perceptual representation of the text on his t-shirt. Analogously, a movie maker need only present an audience with sufficient detail to enable them to recognize the intended content of a scene and to mold their affective, perceptual, and semantic engagement with the characters and events depicted. The processes of object recognition and identification therefore provide a sparsely constructed, but sharply focused, depicitive scaffold against which cognitive processes can be used to fill in the particulars of the content of a shot or scene.

A Short Story About Perception in Everyday Contexts and at the Movies

The cognitivist model for movies that we propose is derived from a diagnostic recognition framework for object identification (Schyns, 1998), a biased competition model for selective attention (Desimone & Duncan, 1995; Pessoa, Kastner, & Ungerleider, 2002), and discussions of visual routines for everyday activities (Hayhoe & Ballard, 2005). Perceptual systems evolved in lock step with bodies in a process driven by the physical laws and instrumental needs of an organism. The net result is that perceptual systems are special purpose systems designed to deliver just what an organism needs to flexibly and efficiently perform a range of everyday tasks in its ordinary environment. Movies function as attentional engines intentionally designed to focus perception on those aspects of the depicitive scaffolding of shots and scenes diagnostic for their narrative content and meaning. The information structure of the movie will, therefore, perhaps counterintuitively, closely match the information structure of real-time experience.

Diagnostic features can be defined as sets of sensory features sufficient to enable an organism to perceptually recognize the identity, shape, location, and affordances of objects and events in the environment. Some views of objects are more informative than others (Schyns, 1998; Palmer, 1999). In our ordinary interaction with clocks we want to know what time it is, for which a view orthogonal to the plane of the clock face is the most informative. A frontal view of a horse, on the other hand, provides very little information about its size and posture. A side view rectifies some of this, but it obscures information about the breadth of the animal across the shoulders. A three-quarter view thereby provides us with the most information about the size, strength, and demeanor of the horse. Of course, if you want to wind the clock, set the alarm, or check the horse’s teeth, a different view would be better. What’s the take-home point? Diagnosticity is a task-specific notion. Altering our behavior, changing
target task, alters the way we categorize objects and events. Changing the category under which we identify a perceptual stimulus can alter the way we assign salience to its parts and features. This, in turn, entails that, as discussed earlier, perceptual systems do not need to generate a global model of the detail of an organism's environment in ordinary contexts. A representation of task-relevant aspects of objects and events that encodes their relative orientation to the organism will do.

A diagnostic recognition framework for object recognition provides a model for how an organism efficiently tunes perception to its behavioral goals. But it also generates a problem. The environment is replete with information. How does an organism flexibly orient its attention to select just that information diagnostic for the current task? One simple means is perceptual salience. Some features in the environment stand out in a crowd simply because they are different (e.g., abrupt movements or changes in luminance intensity). Perceptually salient environmental features are responsible for exogenous, or externally cued, shifts in attention. However, the features necessary for a task are not always (and likely not even often) the most perceptually salient. Biased competition models for selective attention demonstrate that top-down frontoparietal attentional networks bias perception in everyday contexts by priming sensory systems to the expectation of perceptual features diagnostic for the task at hand in the current behavioral context. Feedback from prefrontal, premotor, and parietal areas associated with spatial working memory, object recognition, motor planning, and motor preparation enhances the firing rates of populations of neurons that encode for potential targets at expected locations and inhibits the encoding of potentially distracting information (Kastner, 2004; Schubotz & von Cramon, 2003; Stevens, Fonlupt, Shiffrar, & Decety, 2000). Similar feedback projections from the amygdala, orbitofrontal cortex, and ventromedial prefrontal cortex facilitate the influence of emotional salience on the sensory encoding of diagnostic features (Duncan & Barrett, 2007; Pessoa & Adolphs, 2010). The net result is that perception is naturally tuned to diagnosticity. We perceive just what we need for the current behavioral context. Therefore, biased competition models of attention explain how cue diagnosticity influences perception.

What drives these processes? The goals and expectations of the organism, goals and expectations that are derived from an understanding of the current context and explicit declarative knowledge about the structure and function of object and event types. However, the influence of these variables in perception need not depend on explicit conscious cognitive processes. We live in a stable environment that supports stereotyped behaviors in everyday contexts. This entails that we don't need to generate a global perceptual model of the environment in order to realize our goals. Rather, we can direct attention to those features we need as we need them. We can offload a great deal of the cognitive load of perception to the environment itself. We need only encode enough information to direct our attention to the locations of diagnostic features—or in the vernacular of embodied cognition, the world itself can serve as its own representation (Brooks, 1997). This, in turn, entails that stereotyped visual routines can be developed for directing attention to diagnostic features across stages of ordinary tasks in stable environments. Expert cricket batsmen, for instance, do not track a pitch continuously across space. Instead, they fixate on the release point of the pitch and then shift attention to the point where the ball will bounce in front of them (Land & McLeod, 2000). Likewise, the eye tracks directly to where the hand is going next when we are making tea or sandwiches, often jumping to the next location prior to the actual performance of the current stage of the task (Land & Hayhoe, 2001). These results show that in everyday activities we perceive only what is necessary to support behavior, and our patterns of attention are endogenously driven by automated visual routines that enable us to smoothly track diagnostic features.

What is the consequence of all this for a discussion of movies? The recognition prompts that drive naturalistic pictorial representation can be quite sparse—movie makers can employ stylistic conventions to drive attention to minimal sets of perceptual cues diagnostic for the depictive, narrative, and artistically salient content of a work. Therefore, the intense experiential grip of a movie need not depend on a replete, perceptually rich experience. In the case of movies, the related goals of telling and understanding a story will, in conjunction with constraints set by the medium, govern the selection of what is shown and seen.

**EXOGENOUS CUES: DELIVERING JUST WHAT IS NEEDED**

Selective is a central problem for perceptual systems. The problem is a question of how to filter what is behaviorally salient from the vast array of information present in the environment. Biased competition models for selective attention provide a mechanism that resolves this issue—endogenously enhance the perceptual salience of sensory features to match their cognitive salience while at the same time inhibiting the encoding of any potentially distracting, competing, naturally perceptually salient features. Movie makers solve this problem differently. Theater design, the conventions of theater going, camera movements, lens movements, editing techniques, and the soundtrack are all means to enhance the perceptual salience of features diagnostic for the depictive, narrative, and artistic content of the film. In short, by means of these devices, movies solve the selection process for us. The easiest place to see this is in the theater context. We sit facing a screen, which is scaled to encompass a significant part of the visual field. Typically theater screens are larger than viewers, ranging from 10 feet to 117 feet (the Panasonic IMAX theater in Darling Harbor, Sydney, Australia) in height. Thus, they dominate the visual field with their content. Given that visual acuity falls off significantly as we move away from the central, or foveal, region of the visual field—at an eccentricity of 10 degrees from fixation there is approximately an 80% loss in visual acuity—the vertical and horizontal extent of the screen more than encompasses the region of the visual field within which we can see clearly (Westheimer, 1987). The lights of the theater go down; the soundtrack of the movie comes up; peripheral distractions, including in most cases the din of conversation that would remind us that we are crammed into cramped seats, disappear; and our visual attention is captured. Indeed, the lightened screen in the darkened cavern is all there is to see. These factors focus our attention onto a narrow region of the visual field where we will find all of the diagnostic information we will need to follow the movie narrative for the next 2 hours, a field of focus that doesn't require us to move or reorient our bodies to perceptually survey it.

Not only is the environment of the movie house structured to minimize the cognitive load of engaging movie narratives, but also a significant amount of the cognitive work necessary to perceive and understand a movie has already been done for us in the production process. It has been, so to speak, offloaded to the compositional structure of the movie itself. The basic elements of a movie are shots and the cinematic sequences constructed from them. Shots and sequences are types of pictorial representations, moving pictures that can be used
to depict not just the content, but also the dynamics of an action or event. A shot is a single uninterrupted camera take with no perceptually detectable temporal or spatial discontinuities. Cinematic sequences are composed of a range of shots that present different vantage points on an action, event, or state of affairs for the purpose of narrating a fiction, depicting an environment, communicating a point of view, or shaping a consumer's attitude toward their content. Shots and sequences can therefore be defined as recognitional prompts that present diagnostic information that enables viewers to perceptually recognize their content in much the same way they recognize everyday objects, actions, and events in ordinary contexts. The basic technique of the movie maker in this regard is variable framing, which is used to articulate the contents of shots and sequences by enhancing the perceptual salience of elements within depicted scenes. Variable framing involves the use of discontinuous camera movement (editing), continuous camera movement (pans, tracks, tilts, dollies, etc.), and lens movements (zooms) in order to index, bracket, and scale diagnostic information within a sequence.

Indexing is a means of pointing, a way of gesturing with the camera to indicate where a viewer should be looking or what the viewer should be looking at. Indexing involves pointing the camera at something, thereby communicating, “Look here!” Indexing occurs naturally when the camera is brought closer to its subject by means of a cut, zoom, or camera movement. Pans and tracking shots also do the trick. When the camera pans left or right it shifts the focus of the viewer's attention. If that movement matches, or tracks, the movement of a feature of the scene against its background, we pick that out as the feature to pay attention to. But just pointing a static camera at the relevant feature is sufficient to draw our attention to it. For example, Eisenstein indexes the wheels of the baby carriage teetering on the edge of the step as what is pertinent for our attention at that moment in the “Odessa Steps” sequence of his Potemkin.

Bracketing occurs naturally in all of these cases. As the camera moves toward an object, character, or event, it narrows the field of view, bracketing out everything beyond the edges of the frame as no longer relevant. When, in the “Cigar Store Indian” episode of Seinfeld, the camera cuts in on the gyro sauce from Elaine’s sandwich dripping onto the TV Guide, there is nothing else for the audience to see—everything else has been bracketed out. Similarly, as the camera moves away from the subject of interest, it reframes it in a larger, more inclusive context. Recall the majestic backward-moving crane shot before the intermission of Gone with the Wind; the visual field moves from one wounded body to a myriad, inviting reflection upon the scale of suffering and sadness that war exacts. The scene is reframed, in short, to redirect the object of our thinking. Finally, tracking shots can serve the same purposes. When the camera follows moving objects or characters, it detaches them from the surround, pulling them out of place, isolating them for scrutiny. At the same time, it reframes the object or character by changing its context.

Scaling is a natural consequence of camera movements. Camera movements change the relative size, orientation, and context of those aspects of the environment that they have drawn attention to. We know that an object, like the basement key in Notorious, is of extreme narrative importance because it looms so large on the screen when in close-up. Of course, bracketing and scaling can also focus attention and highlight what the camera doesn't include. For instance, we know that when the camera zooms in on a central character in action/suspense thrillers or horror movies, something critical and "unexpected" is about to befall them, often penetrating our field of vision from off-screen. The close-up brackets our awareness of the movie environment in a way that mimics the scope of the character’s attention. We are carried helplessly through the scene by the character’s unfortunate disregard for his or her surround—a disregard that we recognize leaves them vulnerable to an impending event that will function as a turning point in the movie. This convention causes us to anticipate being startled.

SUPER 8

We can use the train wreck scene from Super 8 to illustrate the uses of variable framing and editing. The movie is about four adolescent boys who are making a zombie movie for a film festival with a Super 8 home movie camera. The boys have grown up together. They are bonded, friends for life. Charles Krznyc, the ringleader of their production, has read that adding a romantic subplot to a movie makes it more interesting. Consequently, he has asked a girl from their class, Alice Dalnard, to join them and play the female supporting role. Of course, it just so happens that she is from the wrong side of the tracks and is willing to sneak away in her dad’s canary yellow muscle car (in the middle of the night) so that they can all drive to the abandoned railroad station on the outskirts of town to film a scene. The main character, Joe Lamb, is quite smitten with the girl. So the introduction of the subplot to the boys’ movie is also the introduction of the movie’s romantic subplot—which is reflected in the fact that Joe is responsible for the makeup on the set and thereby responsible for the presentation of her character (as she develops from a loving wife into a zombie in back) within their production.

A train derailment, accompanied by the usual fantastical explosions and flying debris, occurs while the kids are filming Alice’s scene at the railroad station. At the onset of the disaster the boys run away from the fracas in the direction the train is moving. Alice lingers, watching the train approach. Joe calls out to her but she flies in the opposite direction, toward the oncoming train. Once the commotion settles, the sequence of shots tracks Joe’s wandering path, exploring the wreckage and gathering his friends back into the scene. Alice is conspicuously missing, a point that Charles reinforces by explicitly asking, “Have you guys seen Alice?” At just this point the camera cuts to a close-up of Joe. His eyes focus on something off screen in the foreground. The camera cuts and zooms in on a large piece of twisted metal stained red along its edge. Joe’s sightline and expression prepare our attention and prime our understanding of what follows. The implication is obvious. The sequence settles back to a close-up of Joe staring down at the piece of wreckage. Two of the other boys enter the shot from the left, 10 yards behind him and slightly out of focus, their sightlines visually aimed in the same direction. The camera lingers here a moment, allowing us to bear witness to the putative tragedy … until Alice enters the shot in the background from the right, out of focus. Her movement abruptly contrasts with the stillness of the scene. It is a naturally perceptually salient cue that immediately grabs our attention. As she walks toward Joe, the camera cuts to a sharply focused close-up of her face. She asks, “What’s all that blood?” The camera returns to Joe, who turns and smiles, and then back to Alice, who asks again, “What’s all that blood? Did someone get hurt?” The camera cuts back to a medium shot of Joe from Alice’s perspective. He is standing beside the bloody wreckage. He unhesitatingly leans down and reaches in under the metal, exploring the space with his arm. His body language clearly
tells us he has found something. We recoil and wait with disgusted anticipation as he drags it toward him. He pulls out his makeup kit and, with a giant smile, turns and says, "It's my fake blood. It's fake!" This scene demonstrates the ways inclusive and exclusive bracketing, scaling, and indexing can be used together (along with some narrative flourish, which we will discuss later) to enhance the salience of the elements of the scene. It also shows how dialogue and gaze strategies can be used as exogenous cues to reinforce and enhance the functionality of these devices.

**VISUAL ROUTINES: TUNING PERCEPTION TO THE TASK AT HAND**

The use of variable framing within many (but scarcely all) cinematic sequences approximates the structure and pacing of the visual routines that govern perception in everyday behavior. This is not a novel observation. It was first noted in the silent period by the filmmaker and theorist Vsevolod Pudovkin, who urged that editors build certain scenes around what they would imagine to be the itinerary of the glance of an interested observer of the events being narrated (Pudovkin, 1926). In the ordinary context of everyday activities we do not scan the environment searching for what we need. Rather, evidence suggests that perception is closely coupled with behavior and fine tuned to the environmental features diagnostic for the performance of a task. Hayhoe and her colleagues have demonstrated that we make very few task-irrelevant fixations while performing the common actions associated with tea and sandwich making (Land & Hayhoe, 2001). Participants in these studies make a short initial scan of the room at the start of the task in order to locate salient objects. During this phase participants fixate on task-relevant objects about half of the time. After this initial period, however, participants exhibit very little other orienting, or looking around, behavior. Fewer than 5% of observed fixations were focused on task-irrelevant objects during the performance of the prescribed everyday activities. Further, as discussed earlier, eye movements preceded the motor acts they were associated with by up to half a second, and the focus of attention tended to move on to the subsequent location prior to the completion of a motor act. Finally, although there was variance in the order in which the basic actions, or subroutines, associated with these activities were performed, the gaze strategies within those subroutines generalized across participants. These studies indicate that perception in ordinary contexts is guided by shared internal scripts that produce stereotyped patterns of attention targeted to the requirements of everyday activities. The purpose of these scripts is to provide just the perceptual information needed at just the right time to support current goal-directed behaviors. Dana Ballard has thereby referred to them as *just-in-time strategies* (Ballard, Hayhoe, & Pelz, 1995).

This model for natural vision makes sense. The primary purpose of perceptual systems is to provide information to support our everyday physical and cognitive behavior in a normal environment that has remained, by and large, stable over the duration of the history of the species. It may introspectively feel as if we are natural-born philosophers, reflectively wandering the world in search of meaning, scanning for information salient to our goals. However, the bulk of each day is taken up with commonplace activities like dialing a telephone, operating machinery, writing at a desk, or getting lunch at the cafeteria. A significant proportion of the rest of what we do involves smooth coping with the environment, which involves the same kinds of stereotyped behaviors (e.g., turning door knobs, navigating hallways, and sitting down or getting up from desks or work stations). This entails that the cadence of ordinary conscious experience is, by and large, dictated by the cadence of those patterns of attention associated with those visual routines that govern these everyday activities. Likewise, the camera typically doesn't smoothly track across or actions across the global development of a cinematic sequence. It jumps from one salient feature to the next, mimicking visual routines, building up the content of the depicted action or event out of a set of associated shots (e.g., from a broad view of the wreckage, to a close-up of a twisted piece of metal that was not located within the original tracking shot, to a reverse angle shot of the characters, back to the twisted piece of metal, and then finishing on a close-up of the main character, who is now oriented 180 degrees away from his last position facing his friends).

Of course, there is a diastology here. Visual routines are task-specific internal scripts. They direct attention to just those aspects of the environment an organism needs to interact with in order to realize an internally selected goal. In this regard, visual routines are patterns of attention paired to motor acts necessary for everyday behaviors. However, viewers aren't doing anything at the movies. The diagnosticity of the information presented in shots and sequences is, to a degree it is paired to action at all, paired to the behavior of the characters. Movies are externally imposed attentional scripts that capture a viewer's attention. The viewer is, in a sense, a passive participant whose attention is entrained to the informational structure of the movie, the communicative intentions of the movie maker. Nonetheless, visual routines are automated patterns of attention that seamlessly direct perception across space to just those diagnostic features salient to each stage of a stereotyped behavior. In this regard, the shot/sequence structure of scenes constructed on the interested observer model suggested by Pudovkin is a stand-in for the visual routines that keep our attention fixed to diagnostic features in ordinary contexts— they map to the visual routines that would putatively govern our patterns of attention if we were present in the depicted actions. The frequency of such scenes helps explain the widespread accessibility of movies as well as why they feel so real—sequences constructed on the interested observer model mimic the structure and cadence of those patterns of attention constitutive of the everyday perceptual experiences that they depict (see Smith, Levin, & Cartling, 2012).

Of course, not all shot sequences are modeled on the interested observer model. The cut from the shot of the primeval hurling the bone skyward to the shot of the space station in *2001* is not modeled on anything that an observer, even an ideal observer, could witness. In these cases, the variable framing "leads" our attention to what we need to see to follow the story or to reflect on its significance. Nevertheless, we don't experience this as being coerced because it, too, is based on a sort of just-in-time strategy with regard to the story. We see and come to know just what we need to in order to come to know just what we need to know just when we need to know it. Roughly speaking, changes in camera positions are ordinarily governed by the needs of narrative understanding, by the goal of smoothly and efficiently drawing attention to those aspects of a shot diagnostic for the narrative arc of the story. Rather than burdening the viewers with the task of parsing the action themselves, the movie maker prefocuses their attention in a way that is closely coupled to the instrumental needs of following

---

*4 Despite the seemingly passive character of movie-going experience, there is evidence that motor simulation plays a role in the process of narrative understanding at the movies, in comprehending the actions depicted on the screen (Speer, Reynolds, Swallow, & Zacks, 2009; Zacks, Speer, & Reynolds, 2009).*
the story, the cadence of which apes our ordinary perceptual habits. Variable framing serves that end by, generally, showing us exactly what we need to see just as we need it to carry on with the task of tracking the narrative. In this way, the virtually effortless lucidity of the movies renders their content transparently available to nearly everyone. The goals of moviegoers and the movie makers therefore mesh, and, because perception is naturally coupled to the task demands of our ordinary behavior, this extraordinary experience seems as nothing out of the ordinary, again contributing to the accessibility of movies.

**ENDOGENOUS CUES: CONSTRUCTING A CONTEXT**

It doesn’t take much reflective analysis to notice that there is more to the structure of the sequence from *Super 8* described earlier than the natural perceptual salience of its compositional features. There are two critical variables in the attribution of diagnostically to a perceptual stimulus: the availability of stimulus features and the current goals that frame the context of an organism’s behavior. Movies deliver just the information needed to recover their content—they control the availability of salient stimulus features. Variable framing determines the perceptual trajectory of the unfolding movie narrative, what the viewer sees and the order he or she sees it. However, the viewer isn’t simply passively guided through the depictive trajectory of the spatiotemporally discontinuous shots constitutive of a cinematic sequence. He or she must assimilate what is seen into the larger narrative context in order to recognize what is going on, to interpret and evaluate the story as it unfolds. Furthermore, the novel information presented in shots and sequences functions in much the same way that novel experience functions in ordinary contexts—as developing knowledge that must be accommodated to maintain a coherent and unified model of the story world (Carroll, 2008).

What drives the processes of assimilation and accommodation in our engagement with movies? Loosely speaking, viewers use narrative schema to develop a flexible dynamic model of the unfolding events in the story, which they update and modify on the fly to accommodate novel information presented in the unfolding perceptual trajectory of the movie’s shots and sequences. For instance, Joe’s mother was crushed in an accident at a steel mill just prior to the start of the story in *Super 8*. We (along with Joe) learn from Alice later in the movie that Joe’s mother was covering for Alice’s father at the mill the day she died, and that he had missed work because he was drunk. The intensity of Alice’s question about the bloody twisted steel wreckage retroactively takes on a heightened significance, as does the bonding experience of the train wreck itself, in the context of this new information. The global model that the viewer has constructed for the narrative must be updated, modified to accommodate this shift in our understanding of Alice’s goals, intentions, and perspective.

What is the source of these narrative schema? Just as visual routines are schema that fine tune perception to the spatiotemporal structure of everyday behaviors, folk psychology provides a range of schema, heuristics, and rules of thumb that enable us to interpret and predict the behaviors of fellow human beings in terms of the mental states stereotypically associated with ordinary actions; for example, why did Mr. P stop off at The West End on his way home after long day at work? Feeling beat and worn down, he wanted to unwind, he believed that a game of pool and a beer with some friends would satisfy this desire, and he believed that he could find all three at his local bar. Our understanding of folk psychology and the general causal structure of the world enables us to stitch the information presented in shots and sequences together as inferences to the best explanation about what is going on in the story. These schema are powerful attentional devices movie makers use to produce narrative expectations and to fill in the gaps in a spatiotemporally discontinuous, perceptually sparse narrative. Within this context, movies employ stereotypic narrative strategies in order to provoke and sustain audience attention. Notable among these strategies is erotetic narration (Carroll, 2009).

Erotetic narratives are constructed by generating and answering a hierarchically organized set of questions. These questions set the context that determines the diagnosticity of the elements of individual shots and thereby shape the patterns of attention that define a viewer’s experience at the movies. For instance, *Super 8* opens with a scene depicting the funeral of Joe’s mother. Joe is sitting outside in the snow when a man pulls up in a canary yellow muscle car, takes a wig of a bottle of liquor, and walks into the house. A commotion ensues and Joe’s dad, a local sheriff’s deputy, comes out, clearly upset, with the man in handcuffs. The connection between this event and the funeral event is not explained, nor are we given any clues. So a range of broad narrative questions and attendant expectations immediately come to mind when Alice pulls up in the same car to drive to the location shoot. What was the relationship between Alice’s father and Joe’s mother? Why is Joe’s father upset with Alice’s? How will answers to these questions shape Joe and Alice’s relationship and fuel the dramatic conflict that will drive the story?

Erotetic narrative structure is also used to shape our engagement with more local cinematic sequences. Alice is furious with Charles when she discovers that Joe is coming with them in the car. She calls him “the sheriff’s kid.” He threatens to slip out on the zombie production altogether. This is initially a narratively plausible scenario. It seems unlikely that the tough girl from the other side of town would have agreed to join this group of introverted adolescent boys. Why is she there? Will she really go through with it? However, after she reluctantly agrees to participate, Joe reaches forward from the back seat and offers her a piece of licorice. She accepts it without either hesitation or acknowledgment, signaling both that she is going to stick with the boys and that part of the draw of the group is a curiosity about Joe—a curiosity that is reinforced in the intimacy of their conversation while he does her makeup. All of this information is ready to hand because of our universal natural capacity to use folk psychology to interpret the beliefs and desires of others in stereotyped behavioral contexts.

The final piece of the cognitivist story about the movies involves affective responses to events and behaviors depicted in the movies. Why, for instance, do we recognize that Alice’s sharp reaction to Joe’s presence is a conventional cue to her tact, perhaps as yet unrecognized attraction to the forbidden boy from the other side of the tracks, do we find their unfolding romance compelling? Why don’t we experience it for what it is, a pedestrian stylistic trope? The cognitivist answer, once again, is that the emotional power of movies lies in the efficient way they are used to deliver information and control our attention. The key here is to recognize that affective responses in general, and emotional responses in particular, are evolved means for appraising the behavioral significance of environmental stimuli. We can think of our affective responses to the environment as being divided into two types: involuntary, automatic reflexive responses like the startle response, and more cognitively nuanced emotional responses. Movies employ both as attentional engines, often together. For instance, explosions, sudden movements, and quick changes in the size of objects that mimic visual
looming effects startle us and thereby produce intense autonomic responses that grab our attention. These kinds of effects can, in turn, be used to ramp up the intensity of movie-going experience and prime viewers’ attention to features diagnostic for the unfolding emotional content of the story.

Emotions involve autonomic responses to environmental stimuli that are paired with, and triggered by, appraisals of the potential behavioral significance of objects, agents, events, and actions. There is some disagreement about whether affective appraisals are best defined cognitively, as judgments about the significance of environmental stimuli, or behaviorally, as stereotyped visceromotor responses whose meaning is retroactively interpreted (for a review see Prinz, 2004). However, there is broad agreement that the goal of these appraisals is to categorize the behavioral salience of the relationship between their object and an agent. For instance, the perception of a present threat in the environment may cause the hair on my neck to stand up and produce a chill in my spine, a set of autonomic responses that cause me to freeze defensively in place. Or, the thought of some recent risk or threat may produce a sharp feeling of nausea that temporarily paralyzes my thoughts, causing me to stop and reflect on whether to proceed. The cognitive component of the etiology of these behaviors is so strong that merely imagining the context of an emotionally charged event is often sufficient to induce the autonomic effects associated with an emotional appraisal; for example, recollecting the events surrounding some past injury, say, driving a quarter-inch-wide bit through the palm of your hand during a particularly arduous sculpture installation, can cause one to recoil in mock pain, a stereotyped withdrawal response. In fact, the pull of affective appraisals is so strong that merely imagining these kinds of events can induce their attendant gut reactions (as should be the case for the description of this shop injury).

These facts about the etiology of emotional experience provide a mechanism to explain the intense experiential grip of movies. Variable framing can be used to present information diagnostic for affective appraisals that cognitively triggers contextually appropriate gut reactions appropriate to an emotional response. Emotional appraisals are elicited by general types of situations that are criterially, or categorically, appropriate to them; for example, dangerous situations that are a threat to our safety are criterially appropriate elicitors of the emotional appraisal of fear. Movie makers are able to exploit this mechanism by criterially prefocusing scenes and sequences in which the eliciting factors pertinent to provoking a certain emotion are made salient, typically by means of variable framing. In other words, imagining what variable framing shows us in the local context of a cinematic sequence is sufficient to induce strong emotional reactions in a viewer, which, in turn, influences the way the audience tracks the action.

There are a range of ways these kinds of effects can be enhanced. The pacing of the attentional pattern within a local cinematic sequence can be used to ramp up the perceived tension in a scene, priming the vulnerability of the viewer to information diagnostic for affective appraisals embedded within them. Alternatively, global narrative cues can be used to frame moral expectations about the outcomes of events and the demises of certain characters, expectations that can drive our emotional engagement with a movie narrative. If a movie maker can direct us to perceptually categorize an object or event appropriately, the intended emotion will follow naturally in our movie-going experience. For instance, suspense involves a desire for a just outcome for the protagonist and a set of expectations that define that outcome as improbable within the context of a viewer’s current model for the narrative. Similarly, there are two critical aspects to motion picture horror. First, local narrative and visual cues are used to categorize the behavior and appearance of a character as unnatural, and thereby disgusting. Second, global narrative cues are used to generate the long-term expectation that a negative outcome is highly probable, and likely inevitable, for the protagonist. The result is an intermingling of hopelessness, fear, and revulsion that we delight in experiencing at the movies.

Diagnostics, Attention, and Perception at the Movies

Variable framing is a tool used by movie makers to present diagnostic information sufficient to enable viewers to reconstruct the narrative content of movies from the depictive content of sequences of spatiotemporally discontinuous shots. In this regard movies are attentional engines that are fine-tuned to the shared perceptual capacities of viewers. In this section we flesh out the short story about perception in everyday contexts and at the movies sketched in the sections above (see also Kastner, 2004; Kozbelt & Seeley, 2007; Pessoa et al., 2003; Seeley, 2012). Biased competition models for selective attention identify top-down cortico-thalamic networks that influence the way populations of neurons encode information and communicate at all stages of perceptual processing, from subcortical areas (e.g., the thalamus and superior colliculus in visual processing) to areas associated with higher level perceptual processing (e.g., ventral stream processing areas TE and TEO associated with visual pattern and object recognition). These feedback projections facilitate the implementation of forward models of perceptual change in the environment that enhance the firing rates of populations of neurons that would encode the presence of expected diagnostic features and, as a result, inhibit the firing rates of neurons that encode for perceptual features that are not behaviorally salient in the current context (Moran & Desimone, 1985; Pessoa et al., 2003; Sundberg, Mitchell, & Reynolds, 2009). Biased competition models of selective attention thereby provide a mechanism for fine tuning perception to the goals and expectations of the organism. These models, in turn, suggest a neurophysiological mechanism that can explain how movie makers use variable framing and erotic cues to fine-tune a viewer’s movie-going experience to narratively salient perceptual expectations.

Selective attention facilitates the top-down cognitive control of perception via a cortico-thalamic attentional network. This network includes projections from areas of the dorsolateral prefrontal cortex (dPFC) that are associated with spatial working memory to the frontal eye fields (FEF), an area that is associated with endogenous shifts of visual attention; projections from dPFC to anterior areas of the inferior temporal cortex, TE, and TEO which are each associated with the recognition of complex visual patterns in object recognition (Kastner, 2004); projections from TE and TEO to areas V4 and MT, which are responsible for aligning color, abstract pattern, motion, and depth information from sensory inputs; and projections from dPFC and FEF to the superior colliculus (SC), the lateral geniculate nucleus (LGN), and the pulvinar. The LGN is a primary relay station between the retina and the occipital cortex. Attentional priming to the LGN has been demonstrated to facilitate neural synchronization and enhance the firing rates of populations of neurons that encode the targets of endogenously cued perceptual attention. The pulvinar is a topologically organized structure sensitive to orientation and motion information with connections to all areas of the visual system that is also hypothesized to regulate cortical synchrony (Saillmann
et al., 2012; Pessoa & Adolphs, 2010). The SC is a structure critical to directing eye movements and the cross-modal perceptual integration of visual, auditory, and somatosensory information (Beck & Kastner, 2009). One hypothesized role of the SC is to bring visual, auditory, and somatosensory perceptual information into spatial register with one another in order to direct attention and facilitate the production of a coherent multisensory perceptual representation of the environment (see Grossberg, 1980, 1999; Spence & Driver, 2006; Stein, Stanford, Wallace, Vaughan, & Jiang, 2006).

The SC is reciprocally connected to the auditory system via the inferior colliculus (IC) and the medial geniculate nucleus (MGN) in the auditory thalamus. An analogous processing story can be told about the auditory system. The IC receives top-down input from the auditory cortex, has a large population of auditory-somatosensory multisensory neurons, and plays a role analogous to the SC in a corticofugal auditory attentional network that enhances the encoding of task-salient auditory features and inhibits the perception of local distracters (McLachlan & Wilson, 2010; Winer, 2006). Although there is not space here to expand upon this point, the suggestion is that the model we have proposed can be generalized to include the powerful influence of auditory cues in viewers’ perceptual narrative engagement with movies.

Affective processing also influences perceptual processing via an attentional feedback network that includes projections from the amygdala, orbitofrontal cortex (OFC), ventromedial prefrontal cortex (vmPFC), and anterior cingulate (ACC) to the visual, auditory, and somatosensory systems. The amygdala is reciprocally connected to the visual, auditory, and somatosensory systems and receives inputs from prefrontal areas vmPFC and ACC, which are associated with visceromotor control, or our gut reactions to events in the environment, and the OFC, which is associated with the calculation and categorization of the behavioral significance of objects and events (Duncan & Barrett, 2007). Finally, evidence suggests a motor component to our perceptual responses to narrative sequences. The semantic comprehension of action sequences in movies and texts has been shown to involve activation of those premotor, supplementary motor, and motor areas involved in anticipatory motor preparation for, and the performance of, the depicted actions in ordinary contexts. In ordinary perceptual contexts activation in these areas facilitates tracking and predicting the behaviors of other characters. In narrative contexts these same processes enable movie-goers and readers to use diagnostic cues to model and predict global aspects of depicted behaviors. This broad range of sensory, affective, motor, and attentional processing areas forms an integrated cross-modal circuit that enhances the encoding of features diagnostic for the perception of objects, events, and actions in ordinary perceptual contexts. It suggests a mechanism to explain the cross-modal integration of affective information carried in the soundtrack with the visual narrative. And finally, it explains how sensorimotor cues embedded in cinematic sequences facilitate rich, embodied experiential responses replete with affective dimensions.

Stitching It All Together

So what is the take-away from all of this for our current discussion of the philosophy and neuroscience of motion pictures? A range of tools are used to frame, shape, and drive a viewer’s narrative engagement with movies (e.g., variable framing, editing, erotic cues, and criterial prefocusing on the movie maker’s side and assimilation, accommodation, and emotional appraisals on the viewer’s side). The neurophysiological model for selective attention proposed here provides a mechanism for understanding how viewers stitch the rich narrative experience of the movies together from the sparse collection of raw materials provided by the movie maker. The quick and dirty story is that minimal sets of diagnostic cues are collected and sent downstream for perceptual processing with each glimpse of the environment, every second, every quick shift of attention. This information is matched to declarative knowledge of the structure and function of object and event types and used to generate categorical expectations about both the identity of objects and events in the local environment and the location of further task-salient features, about the structure and location of those objects and parts needed to realize our current goals. Top-down attentional feedback is used, in turn, to prime the visual system to the expectation of these task-salient features at particular locations (see Kosslyn, 1996). This information is used to generate a rough spatial model of salient aspects of the local environment that is held online in spatial working memory and updated over iterations of the feedback loop to accommodate new diagnostic information and the unfolding dynamics of our perceptual world. These same processes are employed offline in spatial reasoning and visual mental imagery tasks to model the perceptual structure and dynamics of scenes, objects, and events. We argue that they are used analogously at the movies to generate perceptual expectations that fill in the gaps and stitch together our narrative understanding of shot and sequence structure.

The interconnectivity between prefrontal areas associated with spatial working memory, motor planning, and motor preparation serves two complementary functions in this general model for perception. First, it facilitates the smooth orienting of our bodies to task-specific perceptual expectations. Second, motor planning and simulation can be used to generate forward models of perceptual change in the environment, to predict how what we perceive will change relative to both our own movements and the movements of other agents. These motor expectations influence what we perceive via the same processes and pathways as perceptual categorization judgments and account for the structure and success of the visual routines that drive our everyday activities. We hypothesize that these expectations about perceptual change enable us to smoothly integrate the spatiotemporally discontinuous shots constitutive of cinematic sequences into coherent local narrative events at the movies.

Affective processes contribute to our perceptual engagement with the environment in analogous ways. The neurophysiological processes that underwrite our capacities to recognize the behavioral salience of objects and events in the world, and govern our gut reactions to them, generate categorical expectations that enhance our perceptual sensitivity to task-salient aspects of our environment. At the movies the directed coupling of valenced anticipation and local perceptual expectations (whether ultimately satisfied or unresolved) is a powerful tool that movie makers use to direct cognitive traffic, shape our understanding of the narrative, and hold us in the grip of their stories.

The trick is that narrative understanding is a cognitive task that involves assimilating sparse sets of spatiotemporally discontinuous local perceptual information (i.e., diagnostic cues) into a coherent and unified, but flexible and dynamic, global narrative model. Our argument is that one reason the experience of these processes seems so natural is that they ape the structure of ordinary perceptual experience, which itself involves integrating sparse sets of diagnostic information into perceptual models and routines that facilitate smooth
Coping in everyday contexts. Therefore, the widespread accessibility and universal intensity of movie-going experience is explained in part by the fact that the compositional structure of movies is fine-tuned to the psychological structures and processes that facilitate smooth coping and social interaction in our ordinary everyday activities.

Caveats and Conclusions

We can imagine someone objecting that our version of cognitivism sounds a lot like a form of film realism—we perceive movies just as we perceive ordinary events in everyday contexts. There is a sense in which he or she would be right. Cognitivism is a form of psychological realism about the way viewers engage with movies. The claim, in contrast to the film language hypothesis, is that there is nothing special or unique about narrative engagement at the movies. Explanations of the widespread intensity and widespread accessibility of movies depend on nothing more than the ordinary psychological processes that drive our everyday interactions with our environment. However, cognitivism is not a form of phenomenological realism about the experience of movies. Viewers clearly recognize that their phenomenological experience at the movies is distinct from everyday life. And in this way it differs substantially from standard theories of film realism that appeal to an illusion thesis about movie-going experience to explain the power of movies.

The results of a recent study by Schwann and Lidirar (2010) present a potential challenge to our claim that recovering the content of shots and sequences requires nothing more than our natural perceptual recognition capacities. In the experiment, native participants, individuals who lived in isolated mountain cabins in southern Turkey and so had no prior exposure to television or cinema, were able to easily understand the referential content of individual film shots but had difficulty describing the content of cinematic sequences constructed from standard editing techniques (e.g., sequences that employed point-of-view shots, establishing shots, and shot/reverse-shot sequences). These results suggest that understanding cinematic sequences may require some degree of learning, some prior familiarity with a range of cinematic conventions. However, there is a caveat. Where the sequences in question depicted individuals in static postures (e.g., two people standing facing one another in a shot/reverse-shot sequence or an individual sitting in a room in a sequence that employed a point-of-view shot), native first-time viewers performed poorly (they provided accurate descriptions of the content of the sequence less than 35% of the time). But, where the sequences depicted individuals engaged in commonplace everyday activities (e.g., preparing tea, cooking a meal, or carrying wood for a fire), the performance of these participants was significantly higher (they provided accurate descriptions on average 56% of the time). These results are not inconsistent with our model. In fact, they may be just what we would have predicted. Our ordinary everyday perceptual activities are situated in environmental and behavioral contexts, each of which constrains the diagnosticity of available perceptual cues. In this light it would be interesting to evaluate whether poor performance with the "static" scenes could be attributed to their abstractness; for example, were there sufficient dynamic postural cues presented in this truncated, contextually isolated set of shots (and was the timing of these cues sufficiently integrated) to clearly indicate the activity the actors were engaged in—in short, were the shots sufficiently structured to clearly provide diagnostic cues to the depicted action? We could easily imagine this to have been the trouble for contextually isolated sequences that depicted an unexpressed, or "hidden," reflective cognitive activity using point-of-view or shot/reverse-shot sequences.

One charge that has been explicitly leveled at the cognitivist program in the study of the moving image is that it is reductionist (Sinnerbrink, 2011). It poses as a theoretical perspective capable of telling us everything we ever wanted to know about movies. It pretends to render endeavors like cinematic interpretation obsolete. But this charge has little merit, at least to the theoretical commitments of this paper. First of all, we do not claim to be addressing every kind of motion picture, but primarily only films, which we understand as the class of mass, mainstream, motion pictures of the fictional narrative variety (including TV of this sort). There are many types of cinematic phenomena that are beyond the scope of the theories propounded in this paper (e.g., the New American Cinema). Moreover, we are not advancing a total theory of even movies. We freely admit that there are aspects of movies about which we have nothing to say. We have not offered a theory of interpreting movies and we do not believe that anything we have said precludes the interpretation of movies. What we have focused on are some of the devices that movies use to enlist and sustain audience attention. These include attentional engines such as variable framing, erotic narrative, and criterial prefocusing. At the same time, we would be the first to say that movie reception involves more than attention and that these other aspects of spectatorship, such as interpretation, demand attention in their own right. Our only point is that provoking and managing the attention of the audience is foundational in the sense that other aspects of spectatorship depend on the audience trained on the screen and on the elements in the array that are relevant to whatever other spectatorial activities the movie maker intends. Thus, we do not think we have presented the whole story in this paper but, at best, a beginning.

References


