Bridging the gap between rich and sparse accounts of vision: A cinematic lens for visual cognition.

Daniel Levin
Vanderbilt University
Department of Psychology and
Human Development
Nashville, TN
1-615-322-1518
daniel.t.levin@vanderbilt.edu

ABSTRACT

In order to understand visual events, we need to process the contents of each view, and then effectively combine information from those views into a coherent understanding of whole scenes and events. On one hypothesis, this process requires intensive visual processing and frequent predictions about upcoming events. On the other extreme is the hypothesis that visual awareness is sparse, and associated with few visual representations and only abstract expectations about upcoming events. In this talk, I will describe some research providing support to the sparse-representation hypothesis by demonstrating that people often remain unaware of inconsistencies in visual properties and event sequence. However, I will also describe research demonstrating that default analyses of events can modulate awareness of visual properties in working memory. Both of these lines of research were inspired by the art of cinema, and I will argue that it is no accident that broadly useful ideas about visual cognition can be mined from the practice of film making which has devoted itself for over a century to creating a visual stimulus that recapitulates the "mental play" that makes meaning of the living visual world.

Categories and Subject Descriptors

H.1.3 [**User/Machine Systems**]

General Terms

Human Factors.

1. INTRODUCTION

In the early 20th century, filmmakers developed a set of editing conventions that produced relatively smooth transitions between shots and allowed different views to be combined into a coherent visual representation of dynamic narrative events. Although some theoretical traditions have hypothesized that these rules work only because viewers have learned the medium-specific skills necessary to "read" film, a more recent tradition has emphasized how the principles of cinema reflect everyday perceptual skills and therefore require relatively little medium-specific learning to operate [1]. If this is true, one might reasonably ask why it is even possible that regular perceptual skills are sufficient for understanding most cinematically-presented events. After all, cinema is different from real-world perception in many ways, not the least of which is the fact that the sequence of views in cinema enters our perceptual system unbidden, associated little of the internal planning normally associated with shifts in attention. In addition, the process of creating and combining shots in cinema is produces a range of between-view inconsistencies such as changes in body positions, properties and spatial arrangement that would presumably never occur in the real world.

These are important objections, and in answering them several interesting hypotheses about both the nature of cinema and about visual cognition suggest themselves. The first hypothesis is often referred to as the sparse encoding hypothesis and it suggests that visual experience is the result of a far more abstract encoding of our visual world than previously thought. The second is that our immediate on-line internal understanding of visual events is a fundamentally heterogeneous mix of perceptual evidence, and internally-generated knowledge and simulations. In this paper, I briefly review evidence supporting both of these hypotheses and argue that exploring these possibilities will require a true interdisciplinary exchange between the arts and sciences.

2. SPARSE VISION

A key question for both film makers and psychologists is how to combine different views of a scene into some understanding of the scene as a whole. On one view, this process might involve a detailed analysis of visual properties in each view and then a search for matching properties which would allow the two views to be aligned much as one might align two overlapping photographs to create a larger panorama. For a while, this was at least part of the explanation for view-combination within psychology, and there even seemed to be evidence of a high-capacity short term visual storage system capable of the job. However, several problems with this hypothesis became apparent. Not only did additional research suggest that the hypothesized short term storage system was not up to the task both in practice and in theory [2], but additional empirical observations suggested that viewers seemed oblivious surprisingly large between-view inconsistencies in visual properties [3]. Some of this latter research, by Dan Simons and myself, documented a phenomenon called "change blindness", and it was partly inspired by film makers observations that between-shot continuity errors often go unnoticed. In our initial experiments we replicated this basic phenomenon by creating a short film packed with continuity errors. Viewers missed a surprisingly large proportion of these errors, both in features that were not necessarily the center of attention (for example, we observed viewers failing to detect that the plates on a table suddenly changed color), and in features that viewers were looking directly at (for example, we substituted the sole actor in a scene for another actor across a simple match-action cut and again found that most viewers failed to detect the change) [4]. These and a host of other findings convinced many psychologists that the visual system does not automatically represent and store visual details unless there is a specific reason for doing so, and, further, that our strong impression that we have a detailed knowledge of our immediate surroundings is strikingly overoptimistic [5]. Instead, people combine views based on much more abstract kinds of consistency such as common foci of attention, consistent goals, and even a logical narrative flow.

Of course, one could argue that all of these experiments were done in various forms of media, ranging from films to pictures to simple object arrays, and that vision operates very differently in the real world, where we clearly must have a much more detailed sense of the surroundings we actually inhabit. However, even very early on our research strongly questioned this idea because we found that it was almost as easy to get people to miss changes in attended real world objects as it was for attended objects in media. Just as participants missed the substitution of one actor for another, we could induce them to miss the sudden substitution of one real-world conversation partner for another (we effected the change by carrying a door between the participant and the first experimenter, leaving one of the experimenters carrying the door behind to finish the conversation) [6].

Findings such as these refute the hypothesis that mediated and real-world vision are fundamentally so different that the latter cannot serve as the basis for the former because mediated vision is inferential while real-world vision is based on a mass of detail. Instead, both are heavily inferential and in important ways based on a sparse set of representations that weave views together by a common narrative [7] shaped by attention to different objects that fit specific roles in specifying meaningful visual events. In other words, even in immediate vision it's the common story that creates a visually integrated understanding of a scene, not the match in visual details across views. This is the essence of a sparse account of visual processing, and even though this idea is, in some ways, controversial, it effectively accounts for the people's striking failures of awareness in both media and the real world.

3. THE HETEROGENEOUS WORKSPACE

A key challenge in following up on a sparse vision account is to explain how the on-line representation of our visual surroundings is formed if it is not based on perceptual detail-matching. There are a number of approaches to this problem but one promising avenue might be to adapt global workspace models of consciousness [8] for their account of how awareness is based on models of the current situation that are flexible combinations of the outputs of a range or brain systems. Thus, the current contents of awareness can be built from a combination of currently attended sense information from vision and other senses, and from internally generated information from memory or other basic cognitive systems. A key feature of these combinations is they can be organized in different ways, and emphasize different kinds of information as the situation calls for it.

Our recent research has emphasized two important elements of this organization. First, it is not consistently temporal, and not consistently predictive. Similar to failures in change detection, we have found that viewers find it very difficult to detect anomalous reversals in the order of events, implying that on-line visual awareness is not associated with any "time-stamp" allowing for easy access to detailed event sequence. Second, we have found that there does seem to be a basic spatial component to the organization of visual event representations. In particular, we have taken inspiration from a filmmakers production heuristic referred to as the 180 degree rule and found that breaking this rule seems to produce a spatial discontinuity that in turn induces participants to check represented visual properties for consistency across views. Combined, these findings suggest that visual events are understood in an abstract heterogeneous workspace that frequently has a basic spatial organization.

4. CONCLUSION

In conclusion, findings such as these demonstrate that we have a perceptual system flexible and intelligent enough not to box itself in to a literal encoding of the visual properties and events, and to allow previous knowledge a powerful role in interpreting events. This is, in large part, what makes a range of mediated experiences possible because some of the differences between mediated vision and real-world vision were never crucial to seeing in the first place. One key question is whether there is a substantive difference between a "good-enough" visual experience in which a medium is fully understandable at some abstract level, and one with maximum emotional impact and realism. On one view, many of the properties that escape awareness and that are unnecessary for some level of event understanding may nonetheless contribute in important ways to enhancing the valence of mediated experiences. Although it is likely that any given visual detail may, for some reason, become part of some specific viewer's narrative of visual awareness, the perceptual differences between cinema and real-world vision are no great barrier to the impact of the stories it tells.

5. ACKNOWLEDGMENTS

Some of this material is based upon work supported by the National Science Foundation under Grant No. 0826701 to DTL.

6. REFERENCES

[1] Smith, T.J., Levin, D.T., & Cutting, J.E. 2012. A window on reality: Perceiving edited moving pictures. *Current Directions in Psychological Science.* 21, 107-113.

[2] Irwin, D. 1991. Information integration across saccadic eye movements. *Cognitive Psychology.* 23, 420-456.

[3] Rensink, R.A., O'Regan, J.K., & Clark, J.J. 1997. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*. 8, 368-373.

[4] Levin, D.T., & Simons, D.J. 1997. Failure to detect changes to attended objects in motion pictures. *Psychonomic Bulletin and Review*. 4, 501-506.

[5] Levin, D.T., & Angelone, B.L. 2008. The visual metacognition questionnaire: A measure of intuitions about vision. *American Journal of Psychology.* 121, 451-472.

[6] Simons, D.J., & Levin, D.T. 1998. Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin and Review. 5*, 644-649.

[7] Dennett, D. 1991. Consciousness Explained. Boston, MA: Little, Brown, and Company.

[8] Baars, B.J., 2002. The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences*. 6, 47-52.