Visual Attention in Edited Dynamical Images

Ulrich Ansorge[1], Shelley Buchinger[2], Christian Valuch[2],
Aniello Raffaele Patrone[2] and Otmar Scherzer[3]
[1]*University of Vienna, Faculty of Psychology, Liebiggasse 4, 1010 Vienna, Austria*
[2]*University of Vienna, Cognitive Science Research Platform, Universittsstr. 7, 1010 Vienna, Austria*
[3]*University of Vienna, Faculty of Mathematics, Oskar Morgenstern-Platz 1, 1090 Vienna, Austria*

Keywords: Attention, Eye Movements, Visual Motion, Video, Editing, Saliency.

Abstract: Edited (or cut) dynamical images are created by changing perspectives in imaging devices, such as videos, or graphical animations. They are abundant in everyday and working life. However little is known about how attention is steered with regard to this material. Here we propose a simple two-step architecture of gaze control for this situation. This model relies on (1) a down-weighting of repeated information contained in optic flow within takes (between cuts), and (2) an up-weighting of repeated information between takes (across cuts). This architecture is both parsimonious and realistic. We outline the evidence speaking for this architecture and also identify the outstanding questions.

1 INTRODUCTION

Our visual world is complex and rich in detail but the human mind has a finite cognitive capacity. This is one of the reasons why humans pick up only a fraction of the visual information from their environment. At each instance in time, humans select only some visual information for purposes such as in-depth recognition, action control, or later retrieval from memory, whereas other visual information is ignored in varying degrees. This fact is called selective visual attention.

One particularly widespread source of visual information is technical dynamic visual displays. These displays depict images of visual motion and are used in computers, mobile telephones, or diverse professional imaging devices (e.g., in devices for medical diagnosis). Importantly, the widespread use of technical dynamic visual displays in human daily life during entertainment (e.g. video), communication (e.g. smart phones), and at work (e.g. computer screens) significantly adds to the visual complexity of our world. An accurate and ecologically valid model of human visual attention is essential for the optimization of technical visual displays, so that relevant information can be displayed in the place and at the right time in order to be effectively and reliably recognized by the user.

One important characteristic of videos and other technical motion images that contrasts with the dynamics of 3-D vision under more natural conditions is the fact that this material is highly edited (or cut). Videos consist of takes and cuts between takes. In this context, takes denote the phases of spatio-temporally continuous image sequences. By contrast, cuts are the spatio-temporal discontinuities by which two different takes (e.g., taken on different days, at different locations, or from different camera angles at the same location) can be temporally juxtaposed at the very same image location. Despite the fact that edited material conveys a substantial part of the visual information that competes for human selective attention, little is known about the way that attention operates in this situation. Specifically, attention research in this domain has almost exclusively focused on the impact of image motion per se (Böhme et al., 2006; Carmi and Itti, 2006; Mital et al., 2013), without paying too much attention to the very different cognitive requirements imposed by extracting information from takes versus cuts. Here, we propose a two-step model in response to this demand. In this model, within takes (between cuts) viewers would attend to novel information and would down-weigh repeated visual input.

In the following, we will develop our arguments for this model. We start with the simplest conceivable bottom-up model, and proceed by a brief discussion of top-down factors as one additional important factor. We then introduce our two-step model as a more realistic and yet parsimonious extension of existing

Ansorge U., Buchinger S., Valuch C., Patrone A. and Scherzer O..
Visual Attention in Edited Dynamical Images.
DOI: 10.5220/0005101901980205
In *Proceedings of the 11th International Conference on Signal Processing and Multimedia Applications* (SIGMAP-2014), pages 198-205
ISBN: 978-989-758-046-8

bottom-up and top-down models. Next, we turn to review the evidence that is in line with our model. Finally, we conclude with a discussion of the outstanding questions.

2 RELATED WORK

To understand how selective visual attention works in humans, one can investigate gaze direction, visual search performance, and visual recognition. The relationship between these three measures will be explained next. To start with, we know that gaze direction is tightly linked to interest, attention, and recognition. Eye movements are an objective index of the direction of visual attention. This assumption is well supported by research on recognition during saccade programming (Deubel and Schneider, 1996).It is therefore not surprising that eye movements provide important cues to the personal intentions and interests of another person.When observing another individual, we use direction of fixation (when the eyes are still), of saccades (when the eyes move quickly from one location to another), or of smooth pursuit eye movements (when the eyes track a moving object in the environment) as a window into the other individual's mind.

Of course, gaze direction is not perfectly aligned with attention and does not always tell us what another person sees (Posner, 1980). For this reason, in attention research, one cannot rely on fixation directions alone. If one wants to understand, where attention is directed, one has to equally draw on conclusions from visual search and visual recognition performance (Treisman and Gelade, 1980).

2.1 The Bottom-Up Model

What is true of attention in general is also true of the so called bottom-up model of visual attention. The bottom-up model is supported by both visual search behavior (Theeuwes, 2010) and eye-tracking (Itti et al., 1998), and its charms lie in its simplicity and parsimony. Bottom-up models rely on one simple principle: "the strength of the visual signal" to explain where humans direct their visual attention. These models disregard different human goals, interests, and other top-down influences, such as prior experiences of an individual, or also task- and situation-specific factors. Instead, bottom-up models define the principles of visual attention in simple objective terms and assume that the focus of attention is fully determined by the characteristics of the momentary visual stimuli in the environment (Frintrop et al., 2010; Itti and Koch, 2001).

2.2 Beyond Bottom-Up Influences

Despite the evidence supporting the bottom-up model, this model is not satisfying because humans do not all look in a task-unspecific way at the same locations (Torralba et al., 2006). But how could individual goals influence visual attention? Top-down models explain this. They emphasize past experiences, goals, intentions, interpretations, and interests of the viewer as predictors of visual attention (Torralba et al., 2006; Wolfe, 1994). Top-down principles can influence seeing and looking in two ways: They either boost the subjectively interesting image features or they deemphasize the subjectively uninteresting image features for the summed salience. Top-down models assign different weights to specific features (Wolfe, 1994) or locations (Torralba et al., 2006). Thus, top-down models are suited for accommodating the influence of subjective interests and goals. They can bridge the gap between model behavior and subjective influences for an improved prediction of eye movements and visual recognition into more realistic predictions of visual attention.

What is lacking so far is a convincing top-down model of visual attention for edited dynamically changing visual displays. Given the fact that humans spend much of their time viewing edited videos (on the Internet, television, or in the cinema), it is unfortunate that even the approaches that tried to model top-down influences mostly operated on static images without considering visual changes over time. Progress in this direction has been made in the form of a surprise-capture or novelty-preference model. Researchers observed that during watching of movies, human attention is captured by surprising or novel visual information (Itti and Baldi, 2009). In the surprise model, stimulus information that repeats over time is deemphasized as an attractor of attention. The surprise model is also parsimonious because it requires just one principle of visual memory of what has been seen in the recent past for an explanation of the creation of goal templates.

However, the surprise model is too rigid. It is incorrect to consider visual feature repetition as always being disadvantageous for the attraction of attention. Many experiments have shown that repeated features attract attention (Bar, 2007; Maljkovic and Nakayama, 1994).

3 TWO-STEP MODEL

We suggest that a two-step model of visual attention offers a realistic description of how attention is allocated in videos and other edited dynamic images (e.g. animated computer graphics, or even medical imaging devices). In the two-step model, surprise capture towards novel information and feature priming towards repeated visual information as the two major top-down principles driving visual attention will take turns as a function of one shared steering variable: the temporal coherence of the optic flow across subsequent images (see Figure 1). With the two-step model we thereby seek to overcome existing limitations of (1) bottom-up models that fail to account for inter-individual variability of visual attention, (2) too rigid forms of top-down models of attention that incorporate only one of the two top-down principles, and (3) models that fail to consider the specificities of edited dynamically changing visual images at all. This two-step model is based on empirical observations. It also allows deriving new testable hypotheses that can be investigated with the help of psychological experiments.

To start with, attraction of attention by repeated features (as in feature priming) conflicts with the finding of Itti and Baldi (2009) that repeated features do not attract attention. Attraction of attention by feature repetition can, however, be reconciled with the findings of Itti and Baldi by the two-step model. Itti and Baldi based their conclusions on gaze directions recorded during the viewing of edited video clips and video games. How this could have masked repetition priming across cuts can be understood if one takes into account the specificities of the high temporal resolution of the surprise model that was set to the level of single frames. For each frame, a prior and a new probability distribution were computed and their difference was tested for its potential to attract the eyes. This resulted in a higher number of model tests between cuts (or within takes) of the videos than model tests across cuts (or between takes), even in the highly edited video clips with relatively many cuts. Between-cuts events encompassed 30 frames/second because monitor frequency was set to 60.27 Hz (and assumed that videos were displayed in half frames). However, by definition, each across-cut event consisted of only two frames. Therefore, between-cuts events by far outweighed across-cut events in the test of the model of Itti and Baldi (2009).

Importantly, between cuts (or within takes), the correlation between successive feature or stimulus positions is high, whereas across cuts (or between takes) it is lower. To understand this, think firstly of an example of a take (i.e., a between-cuts event), such as the filming of a moving object in front of a static background. Here the background objects and locations are correlated for all frames of the take. In fact, they would be the same (see Figure 2, for a related example). Now secondly think of what happens across a cut (or between takes). Here, the correlation between successive features or stimulus positions must be lower, simply because of occasional cutting between takes of completely different scenes (or at least different camera angles within the same scene). With temporal juxtaposition of different scenes by a filmic cut, no stimulus contained in the take preceding the cut needs to be repeated after the cut. Basically, this low take-by-take correlation across cuts in videos is exactly what corresponds best to the conditions of the experiments demonstrating feature priming: In psychological experiments, a low correlation between positions and even colors of relevant to-be searched-for target stimuli between trials has been the way to prevent anticipation of target positions and target features (Maljkovic and Nakayama, 1994)). This low correlation corresponds much better to effects across cuts. Basically, in our two-step model we will therefore assume that across cuts, the surprise model of attention would be falsified and a feature priming model would be confirmed, whereas between cuts a preference for novel information holds.

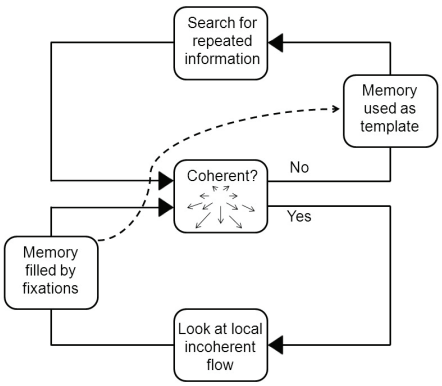

Figure 1: According to the model, within takes the human gaze is steered toward novel information. This mode is supported by the presence of temporally coherent global optic flow (see center of Figure) and an attraction to novelty is achieved by down-weighting global optic flow and up-weighting local incoherent flow for the selection of gaze directions because per definition, the information contained in the global flow field relates present to past information whereas local incoherencies form new features themselves and are diagnostic of the appearance of new objects in the visual field. The situation changes if a cut is encountered. Cuts are signaled by incoherencies of the global flow field. In this situation, the human gaze is steered towards repeated information. For further information, refer to the text.

Figure 2: An image from a sequence of a man shutting the back of his car on the left and a schematic representation of the regions of the highest movement (in black) on the right. As compared to the coherent null vector of optic flow in the background, the optic flow of the moving man would be less coherent and, within a take, should capture human attention.

The two-step model comprises three components: one spatially organized representation of the visual image as its input and two internal top-down representations of visual features. The input representation is the same as in standard bottom-up models (Itti and Koch, 2001). The two alternative internal top-down representations of the two-step model are (1) search templates of scene- or take-specific object-feature matrices that a viewer can retrieve from visual memory, and (2) a track record of the temporal coherence of the optic flow within the image that the viewer applies online while watching a video.

The two-step model's visual memory contains representations of visual feature combinations (e.g. edge representations) for objects and for scenes (or takes). If such a representation is retrieved, this memory representation can be used as a template to upweigh repeated feature combinations as relevant during visual search. This conception of a retrieved search template is very similar to that of other top-down models of attention (Wolfe, 1994; Zelinsky, 2008). In contrast to past feature-search template models, however, as in the surprise model, in the two-step model the content of the visual memory will be empirically specified: What a particular person looks at is stored in visual memory (Maxcey-Richard and Hollingworth, 2013). The two-step model thus uses gaze direction for segmentation and stores objects as a vector of visual features at a fixated position, and each scene or take within a video as a matrix of the vectors of the looked-at objects within a take. Each take-specific matrix will be concluded when a minimum of the temporal coherence of optic flow indicates a change of the scene (see below), and matrices will be successively stored in the order of their storage until a capacity limit of visual working memory has been reached (Luck and Vogel, 1997). In this manner, the two-step model adapts to interindividual variation of looking preferences and keeps track of them, without having to make additional assumptions.

Related but operating on a different time scale, for the two-step model optic flow will be continually calculated as a mathematical function that connects one and the same individual features or objects at subsequent locations in space and time by one joint spatio-temporal transformation rule that is characteristic of the change of the larger part of the image for a minimal duration (Patrone, 2014). Moreover, the temporal coherence of the optic flow will be continuously tracked. We calculate the temporal coherence of optic flow as the similarity of the optic flow across time. In the two-step model, an increasing temporal coherence of optic flow will thus be reflected in a descending differential function. This coherence signal can be topographically represented in image coordinates and directly feeds into one visual filter downweighing those image areas characterized by the temporally coherent optic flow. In this way, the two-step model instantiates the surprise-capture principle and filters out the repeated visual features proportional to the duration and area of uniform optic-flow (see Figure 2).

By contrast to this, the local minima of the coherence of optic flow (or the maxima of the differential function) are used as signals indicating cuts that trigger the retrieval of a search template, and the resultant up-weighing of the repeated features of the image representation resembling the search template.

The two-step model is more realistic than the surprise model because it incorporates feature priming of attention, too. Yet, the two-step model is parsimonious because it couples the two top-down principles of attention to the same shared steering value of optic flow coherence, and, as in the surprise model, most of two-step models free parameters (the content of the visual memory) will not be arbitrarily chosen or have to be specified by task instructions as in standard top-down models (e.g., (Najemnik and Geisler, 2005)) but will be specified on the basis of empirical observation (i.e., will be measured as the feature values at fixated positions).

4 EVIDENCE

4.1 Weighting Coherent Optic Flow

The surprise-capture principle outperforms the bottom-up model when predicting fixations within animated video games and movies (Itti and Baldi, 2009). According to the two-step model, this surprise-capture effect reflects the suppression of coherent optic flow. Optic flow denotes the global commonality or unifying mathematical rule of the global visual motion signal across the image that is frequently due to the cameras (or the observers) self-motion. Optic flow is tied to visual feature repetition because across time, like other types of visual motion, too, optic flow reflects a track record of repeated features and objects found at different places.

In line with the assumed down-weighting of coherent optic flow, visual search for a stationary object is facilitated if it is presented in an optic flow field as compared to its presentation among randomly moving distractors (Royden et al., 2001). Likewise, objects moving relative to the flow field pop out from the background (Rushton et al., 2007). A tendency to discard optic flow as a function of its coherence over time and space in dynamic visual scenes also accounts for many instances of attention towards human action in general (Hasson et al., 2004) and human faces in particular (Foulsham et al., 2010). In these situations, actions and facial movements are defined by local motion patterns that have regularities differing from the larger background's coherent flow field. Equally in line with and more instructive for the present hypothesis are the cases in which one motion singleton among coherently moving distractors captures human attention (Abrams and Christ, 2003; Becker and Horstmann, 2011). To perform further analysis in this direction we are developing a decomposition procedure of the motion in dynamic image sequences. For example, on the web-page `http://www.csc.univie.ac.at/index.php?page=visualattention` a movie presenting the projection of a cube moving over an oscillating background can be viewed. The optical flow computed between the sixteenth and the seventeenth frame of the sequence is visualized in Figure 3. The motion can be decomposed in global movement of the cube depicted in $U1$ and in the background movement in $U2$.

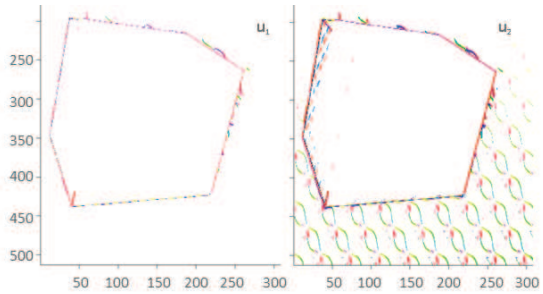

Figure 3: Flow visualization of a dynamic image sequence showing a projection of a cube moving over an oscillating background. $U1$ and $U2$ depict the global and the background movement respectively (`http://www.csc.univie.ac.at/index.php?page=visualattention`).

4.2 Templates Combining Features

Selectively attending to the relevant visual features for directing the eyes and for visual recognition is one way by which humans select visual information in a top-down fashion (Wolfe, 1994). For example, informing the participants prior to a computer experiment about the color of a relevant searched-for target helps the participants in setting up a goal template representation to find the relevantly colored target object and ignoring irrelevantly colored distractors (e.g., to find red berries in green foliage during foraging; (Duncan and Humphreys, 1989). Equally important and well established is the human ability to selectively look-for particular visual shapes or for specific combinations of shapes and colors (Treisman and Gelade, 1980). In this way human viewers could also search for landmarks that they have seen in the past to reorient after a cinematic cut, and to decide whether a visual scene continues or has changed.

In line with this assumption, participants learn to adjust the search templates to the visual search displays that they have seen in the past. During contextual cueing, for example, participants benefit from the repetition of specific search displays later in a visual search experiment (Brooks et al., 2010). Similar advantages have been demonstrated in the context of visual recognition under more natural conditions, with static photographs of natural scenes (Maxcey-Richard and Hollingworth, 2013; Valuch et al., 2013).

In the study of (Valuch et al., 2013), for example, participants first viewed a variety of photographs for later recognition of the learned photographs among novel pictures. Critically, during recognition participants only saw cutouts from scene images. Cutouts from the learned scenes were either from a previously fixated area(see Figure 4) or they were from an at least equally salient non-fixated area of the learned images(see Figure 5).

Figure 4: Cutouts from old images were selected contingent upon the participants gaze pattern. Old/fixated cutouts showed the location of longest fixation. Old/control cutouts showed a nonfixated but highly salient location. Copyright by AVRO (Valuch et al., 2013).

Figure 5: Cutouts from new images showed highly salient scene regions or were randomly chosen. Copyright by AVRO (Valuch et al., 2013).

In line with an active role of fixations for encoding and successful recognition, the participants only recognized cutouts that they fixated during learning. In contrast, the participants were unable to recognize cutouts showing areas that were not fixated during learning with better than chance accuracy (see Figure 6).

4.3 Reorienting After Cinematic Cuts

We consider re-orienting between subsequent visual images as one of the most fundamental tasks for the human viewer. Under ecological conditions, orienting is required in new environments, as well as when time has passed between successive explorations of known environments. During the viewing of edited videos, orienting is required to make sense of temporally juxtaposed images with a low correlation of objects or their locations. The latter situation is typical of all technical imaging devices for dynamically changing visual images. Think of video cuts in which the image before the cut does not have to bear any resemblance to the image after the cut.

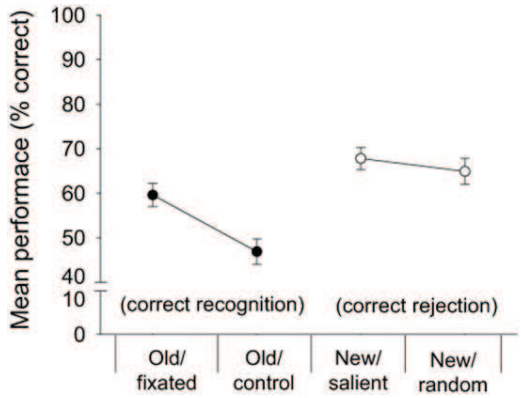

Figure 6: Rate of correct responses in percent as a function of cutout type in the transfer block. Copyright by AVRO (Valuch et al., 2013).

In line with the assumed role of repetition priming on eye movements, Valuch et al. observed that participants preferentially looked at videos baring a high similarity of pre-cut and post-cut images (Valuch et al., 2014). These authors used two videos presented side by side and asked participants to keep their eyes on only one of the videos. Critically, during two kinds of cuts, the images could switch positions: cuts with a high pre- and post-cut feature similarity and cuts with a lower pre- and post-cut feature similarity. For example, participants were asked to look at a ski video and to quickly saccade to the ski video if the video switched from the left to the right side. In this situation, the participants showed a clear preference to look at the more similar images. Saccade latency was much lower in the similar than in the dissimilar condition (see Figure 7).

Repetition priming would also explain why participants fail to notice so-called matching cuts. Participants fail to register matching cuts, such as cuts within actions (with an action starting before the cut and being continued after the cut), as compared to non-matching cuts from one scene to a different scene (Smith et al., 2012). This is because with matching cuts, that is, cuts within the same scene, the overall changes in visual image features between two images are smaller than with cuts that connect two different scenes (Cutting et al., 2012).

4.4 Repeated vs. Novel Features

When researchers rearranged an otherwise coherent take by cutting it and rearranging the take into a new and incoherent temporal sequence, the reliability of the gaze pattern was drastically reduced (Wang et al., 2012). These authors argued that their participants kept track of objects within takes and reset this search tendency after a cut. This interpretation is in line with our view that participants apply different strategies within than between takes.

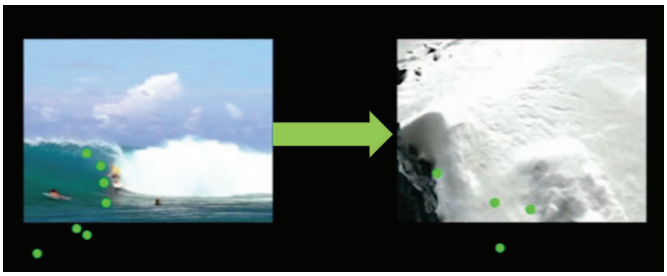

Figure 7: Two videos showing different content were presented side by side and participants were asked to follow only one content with their eyes. Videos could switch position during two kinds of cuts: (1) high and (2) low pre- and post-cut feature similarity cuts. The participants showed a clear preference to look at the more similar images.

5 OPEN QUESTIONS AND POTENTIAL APPLICATIONS

Regarding our two-step architecture many open questions remain. Most critically, it is unclear whether the coherence of optic flow is indeed down-weighted for attention. To address this question, one would need to correlate the decomposition of optical flow (Abhau et al., 2009) into bounded variation and an oscillating component with viewing behavior in natural images. One would also have to test whether changes for novel versus repeated features are characteristic of phases of low global coherence of the flow pattern.

Another open question concerns the impact of top-down search templates for features in natural scenes. This influence is relatively uncertain. Most of the evidence for the use of color during the top-down search for targets stems from laboratory experiments with monochromatic stimuli (Burnham, 2007). This is very different from the situation with more natural images, such as movies, where each color stimulus is polychromatic and consists of a spectrum of colors. In addition, a lot more questions than answers arise with regard to the storage and usage of different take-specific topdown templates.

Among the open questions, the potential applications of the model are maybe the most interesting ones. The model should be useful for improving the prediction of visual attention in more applied contexts, such as clinical diagnosis based on visual motion (e.g., in ultrasound imaging), QoE (quality of experience) assessment and videos coding in entertainment videos.

In medical imaging, much as with cuts, the optical flow of an image sequence can be interrupted by noise or by changes of perspective. For example in case of angiography, a new perspective of the vessels can be suddenly shown.

Also, due to a lack of contact between imaging devices and body (e.g., during ultrasound diagnosis), noise or blank screens can interrupt medical image sequences. These examples illustrate that the two-step model is applicable to medical imaging and that it captures a new angle on these problems. During medical imaging, pervasive eye-tracking could be used to extract visual feature vectors at the looked-at image positions. After an interruption of the imaging sequence, these vectors could then be convolved with post-interruption images for a highlighting of those regions baring the closest resemblance with the input extracted before the interruption. Likewise, in the area of video coding and compression, scene cuts represent an important challenge.

6 CONCLUSION AND COMPARATIVE EVALUATION

With a simple two-step model of down-weighting redundant information contained in optic flow versus up-weighting repeated information contained in two images divided by a cut, we proposed a framework for studying attention in edited dynamic images. This model is very parsimonious because it does not require many assumptions and it can be empirically falsified. In comparison to a bottom-up model the two-step model is able to accommodate inter-individual viewing differences but has more free parameters and is therefore less economical. In comparison to existing top-down models the two-step model takes the particularities of visual dynamics into account and used empirical observations to specify many of its free parameters. Although this model nicely explains a variety of different findings, future studies need to address many outstanding questions concerning the model.

REFERENCES

Abhau, J., Belhachmi, Z., and Scherzer, O. (2009). On a decomposition model for optical flow. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 126–139. Springer.

Abrams, R. A. and Christ, S. E. (2003). Motion onset captures attention. *Psychological Science*, 14(5):427–432.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in cognitive sciences*, 11(7):280–289.

Becker, S. I. and Horstmann, G. (2011). Novelty and saliency in attentional capture by unannounced motion singletons. *Acta psychologica*, 136(3):290–299.

Böhme, M., Dorr, M., Krause, C., Martinetz, T., and Barth, E. (2006). Eye movement predictions on natural videos. *Neurocomputing*, 69(16–18):1996–2004.

Brooks, D. I., Rasmussen, I. P., and Hollingworth, A. (2010). The nesting of search contexts within natural scenes: Evidence from contextual cuing. *Journal of Experimental Psychology: Human Perception and Performance*, 36(6):1406.

Burnham, B. R. (2007). Displaywide visual features associated with a search displays appearance can mediate attentional capture. *Psychonomic Bulletin & Review*, 14(3):392–422.

Carmi, R. and Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research*, 46(26):4333 – 4345.

Cutting, J. E., Brunick, K. L., and Candan, A. (2012). Perceiving event dynamics and parsing hollywood films. *Journal of experimental psychology: human perception and performance*, 38(6):1476.

Deubel, H. and Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, 36(12):1827 – 1837.

Duncan, J. and Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological review*, 96(3):433.

Foulsham, T., Cheng, J. T., Tracy, J. L., Henrich, J., and Kingstone, A. (2010). Gaze allocation in a dynamic situation: Effects of social status and speaking. *Cognition*, 117(3):319–331.

Frintrop, S., Rome, E., and Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, 7(1):6.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *science*, 303(5664):1634–1640.

Itti, L. and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203.

Itti, L., Koch, C., Niebur, E., et al. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259.

Luck, S. J. and Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657):279–281.

Maljkovic, V. and Nakayama, K. (1994). Priming of pop-out: I. role of features. *Memory & cognition*, 22(6):657–672.

Maxcey-Richard, A. M. and Hollingworth, A. (2013). The strategic retention of task-relevant objects in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(3):760.

Mital, P., Smith, T. J., Luke, S., and Henderson, J. (2013). Do low-level visual features have a causal influence on gaze during dynamic scene viewing? *Journal of Vision*, 13(9):144–144.

Najemnik, J. and Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391.

Patrone, A. R. (2014). Optical flow decomposition with time regularization. In *Conference on IMAGING SCIENCE*.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32:3–25–.

Royden, C. S., Wolfe, J. M., and Klempen, N. (2001). Visual search asymmetries in motion and optic flow fields. *Perception & Psychophysics*, 63(3):436–444.

Rushton, S. K., Bradshaw, M. F., and Warren, P. A. (2007). The pop out of scene-relative object movement against retinal motion due to self-movement. *Cognition*, 105(1):237–245.

Smith, T. J., Levin, D., and Cutting, J. E. (2012). A window on reality perceiving edited moving images. *Current Directions in Psychological Science*, 21(2):107–113.

Theeuwes, J. (2010). Top–down and bottom–up control of visual selection. *Acta psychologica*, 135(2):77–99.

Torralba, A., Oliva, A., Castelhano, M. S., and Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766.

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136.

Valuch, C., Ansorge, U., Buchinger, S., Patrone, A. R., and Scherzer, O. (2014). The effect of cinematic cuts on human attention. In *TVX*, pages 119–122.

Valuch, C., Becker, S. I., and Ansorge, U. (2013). Priming of fixations during recognition of natural scenes. *J Vis*, 13(3).

Wang, H. X., Freeman, J., Merriam, E. P., Hasson, U., and Heeger, D. J. (2012). Temporal eye movement strategies during naturalistic viewing. *Journal of vision*, 12(1):16.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological review*, 115(4):787.