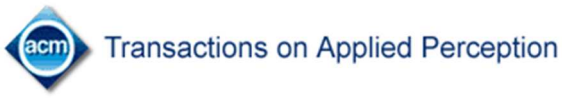

Visual Attention in Edited Dynamical Images

Journal:	*Transactions on Applied Perception*
Manuscript ID:	Draft
Manuscript Type:	Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Ansorge, Ulrich; University of Vienna, Faculty of Psychology Buchinger, Shelley; University of Vienna, Patrone, Aniello; University of Vienna, Valuch, Christian; University of Vienna, Scherzer, Otmar; University of Vienna,
Computing Classification Systems:	A.1 [Introductory and Survey], I.2.10 [Vision and Scene Understanding], I.4 [Image Processing and Computer Vision], I.6.5 [Model Development]

Visual Attention in Edited Dynamical Images

ULRICH ANSORGE, SHELLEY BUCHINGER, ANIELLO PATRONE, CHRISTIAN VALUCH AND OTMAR SCHERZER, University of Vienna

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.

Categories and Subject Descriptors: **A.1** [Introductory and Survey]: **I.2.10** [Vision and Scene Understanding]: **I.4** [Image Processing and Computer Vision]: **I.6.5** [Model Development]

General Terms: Human Factors, Algorithms

Additional Key Words and Phrases: Attention, Eye Movements, Visual Motion, Video, Editing, Saliency

ACM Reference Format:

Ansorge, U., Buchinger, S., Patrone, A., Valuch, C., and Scherzer, O. 2013. Visual Attention in Edited Dynamical Images. ACM Trans. Appl. Percept. ***, ***, Article *** (***), 13 pages.

DOI***

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. Selective visual attention has a central role for the control and execution of thought and action. Just consider the everyday task of reading, which you are performing right in this moment: Different areas of text need to be selected in a serial manner to execute the next saccade (jumping eye movement) to a new word. Similarly, all of our everyday actions require selective visual attention in order to be effectively performed. Because of its importance, understanding human attention is a crucial prerequisite for diverse areas of Cognitive Sciences, especially Vision Science and related disciplines, such as Psychology, Biology, and Image Processing and Computer Vision, to mention a few of them.

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.

This work is supported by the Wiener Wissenschafts-, Forschungs- und Technologiefonds (WWTF) grant number CS-11-009, awarded to Ulrich Ansorge, Shelley Buchinger and Otmar Scherzer.

Author's address: U. Ansorge, University of Vienna, Faculty of Psychology, Liebiggasse 5, 1010 Vienna, Austria; email: ulrich.ansorge@univie.ac.at; S. Buchinger, University of Vienna, Faculty of Computer Science, Lenaugasse 2/8, 1080 Vienna, Austria; email: shelley.buchinger@univie.ac.at; A. Patrone, University of Vienna, Computational Science Center, Oskar Morgenstern-Platz 1, 1090 Vienna, Austria; email: aniello.patrone@univie.ac.at; C. Valuch, University of Vienna, Faculty of Psychology, Liebiggasse 5, 1010 Vienna, Austria; email: christian.valuch@univie.ac.at; O. Scherzer, University of Vienna, Computational Science Center, Oskar Morgenstern-Platz 1, 1090 Vienna, Austria; email: otmar.scherzer@univie.ac.at

DOI***

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, Dorr, Krause, Martinetz, & Barth, 2006; Carmi & Itti, 2006; Itti, 2005; Mital, Smith, Luke, & Henderson, 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. As we will explain in the present paper, we believe that much of this repeated information is contained in optic flow because the optic flow field by definition contains redundant visual input (Horn & Schunck, 1981; Koenderink, 1986; Lee & Kalmus, 1980). By contrast, across takes (i.e., directly after a cut), the viewer's interest should be directed to repeated rather than to novel visual information because visual information conveyed after a cut can be entirely unexpected for the viewer, and only searching for repeated information helps to efficiently re-orient within the new scene. With each cut, the viewer needs to decide whether or not the images after the cut continue or discontinue the situation as seen before the cut (cf. Hochberg & Brooks, 1996, who described this in connection with film perception). The two principles of preferences for novelty and for repeated information take turns in situations other than viewing videos (or films), too, such as when flipping from one web page to another in an online application, or when jumping back and forth between different camera perspectives in a medical imaging device.

In the following, we will develop our arguments for this model. We start with the simplest conceivable bottom-up model. We then introduce our two-step model as a more realistic and yet parsimonious extension of the bottom-up model. Next, we turn to review the evidence that is in line with our model. Finally, we conclude with a discussion of the outstanding questions.

1.1 The Bottom-Up Model

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: It has been repeatedly shown that programming a saccade to a given position *x* within a visual display facilitates recognition at this location *x* but that at the same time, participants cannot covertly attend to a different position than *x* to recognize a visual input at this alternative location (Deubel & Schneider, 1996; Kowler, Anderson, Dosher, & Blaser, 1995). Apart from that, research showed that when complex scenes have to be recognized from photographs, only visual information that was fixated during encoding is reliably encoded into memory, and that the same locations, or objects need to be re-fixated for a later successful recognition (e.g., Valuch, Becker, & Ansorge, 2013). It is therefore not surprising that eye movements provide important cues to the personal intentions and interests of another person: We readily judge others' foci of attention, and their present goals by where they look (Baron-Cohen, 1997). 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 example, recent evidence has shown that participants are able to attend to a position away from the next saccade location if this location can be selected by the same relevant feature (here: a specific color) as was present at the next saccade location (Born, Ansorge, & Kerzel, 2012, 2013). 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 (e.g., Treisman & Gelade, 1980; Wolfe & Horowitz, 2004). In this context, visual search denotes an experimental task that requires searching for a pre-defined target among distractors. The efficiency of visual search is assessed on the basis of search functions. Search functions give search times of successful target searches as a function of the number of distractors in the display. Shallow search functions indicate efficient search, whereas search functions increasing with the number of distractors indicate inefficient search.

What is true of attention in general is also true of the bottom-up model of visual attention. The bottom-up model is supported by both visual search behavior (Theeuwes, 1992; Theeuwes, 2010) and eye-tracking (Itti, Koch, & Niebur,

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 (for an overview, see e.g. Frintrop, Rome, & Christensen, 2010; Itti & Koch, 2001). For instance, many models use local visual feature contrast in the environment as the major property for predicting of fixation directions (see Figure 1). For this approach, in a visual image, per each pixel's 1° surround, for example, the standard deviation of the measured red-green difference is computed. This yields a spatial map of contrasts across the image, a step repeated for different features (e.g., the blue-yellow difference, orientation, and luminance). The resulting local feature contrasts can then be summed, separately for each pixel, and the summed and normalized contrast represented in one spatial salience map. In this salience map, each pixel's summed feature contrast can be compared with all other pixel contrasts. The image areas with the highest summed contrasts are the most salient areas. These image regions attract gaze, one by one, in the order of their salience values from high to low (Itti, Koch, & Niebur, 1989).

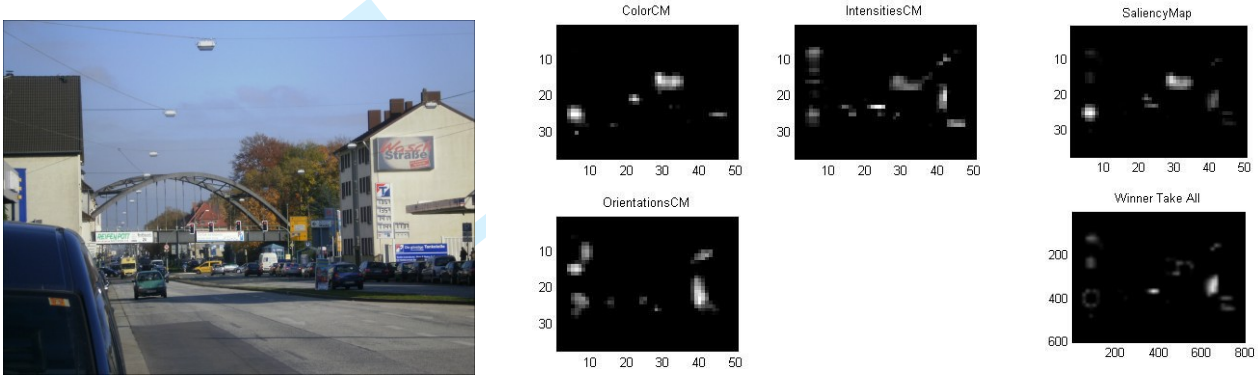

Fig. 1: An image on the left and its feature contrast maps in the middle panel, of which the upper left depicts contrasts of color, the lower left orientation, and the upper right intensity (luminance). In the top right panel, one can see the summed contrasts in the salience. In the bottom right panel, a winner-take-all map of activities shows how likely it is that gaze is attracted towards an image region. Strengths of contrasts and salience increase from dark to light in the middle and right panels. All computations after Walter and Koch (2006).

Also, return of attention or of the gaze to a previously attended-to or fixated position is inhibited by an additional mechanism – inhibition of return (IOR). IOR denotes the reluctance to return to a previously inspected or attended-to image region (Klein, 2000; Posner & Cohen, 1984). It seems to be a primitive kind of memory that supports foraging where the revisiting of previously scanned locations would be non-beneficial (Z. G. Wang & Klein, 2010). Early evidence suggested that IOR might be restricted to the dynamic changes of features, such as abrupt onsets (Gibson & Amelio, 2000) but more recent findings suggested that IOR can also be found for non-changing features, like colors (Ansorge, Priess, & Kerzel, 2013, Priess, Born, & Ansorge, 2012).

1.2 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 (Buswell, 1935; Torralba, Castelhano, & Henderson, 2006; Yarbus, 1967). 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 (e.g., 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 (cf. Wolfe, 1994) or locations (cf. Torralba et al., 2004). If searching for red berries, the modeled observer could, for example, instantiate a goal template with an increased weight of the relevant red-green contrast to a sum of weighted contrasts in the salience map. 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, or the role of prior visual experiences for the current guidance of attention. 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 (cf. Itti & Baldi, 2009). In the surprise model, stimulus information that repeats over time is deemphasized as an attractor of attention. Computationally, past information is represented as a prior probability distribution of visual features. New information is represented as such a probability distribution, too. Attention is then attracted to the most informative image regions. In the surprise model, these are the regions with the largest differences between prior and new probability distributions. The surprise model explains how top-down templates are specified (by past experiences) and it accounts for subjective influences by using the viewers' past experiences as predictors of their gaze direction. 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. In general, a preference for novel stimuli, as proposed by the surprise model, is also an important pre-requisite for the buildup of memory based on ecologically relevant sensory input (cf. Fletcher et al., 2001; Ranganath & Rainer, 2003).

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 (cf. Bar, 2007; Maljkovic & Nakayama, 1994). For feature priming effects on attention, mere feature similarity, resemblance, or relatedness is sufficient. Take the example of the study of Maljkovic and Nakayama. These authors presented their participants with three stimuli in every display, one of which was a so-called color singleton – that is, this stimulus had a unique color (e.g., red) differing from that of the two other stimuli that shared colors (e.g., were both green). Participants had to search for this singleton in each display and reported its shape. Crucially, under these conditions, inter-trial priming was observed: If the color of the singleton repeated from trial to trial, it was easier for the participants to find the target. This was reflected in shorter search times in inter-trial primed trials than in non-primed trials in which the color of the target in a preceding trial *n-1* was different from that of the target in the following trial *n*. Importantly, such priming effects are very robust. For example, initial observations suggested that inter-trial priming could be weak or absent where the salience of the target within a display is strong (Meeter & Olivers, 2005). However, research has demonstrated that this is not the case and that inter-trial priming for the first saccades can be observed, even if the salience of the targets within a display increases (Becker & Ansorge, 2013).

2. A TWO-STEP MODEL OF ATTENTION FOR EDITED DYNAMIC INPUT

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 2). 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 dynamically changing visual displays and the high temporal resolution of the surprise model. The temporal resolution of the model test 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).

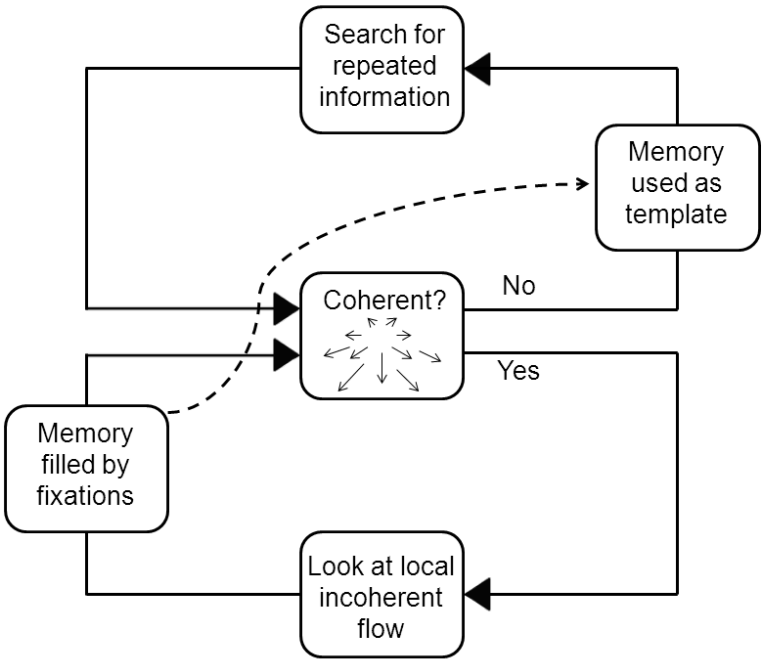

Fig. 2. According to the model, within takes the human gaze is steered toward novel information. This mode is supported by the presence of temporally coherent global optic flow (see center of Figure) and an attraction to novelty is achieved by down-weighting global optic flow and up-weighting local incoherent 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.

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 3, 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 (cf. Maljkovic & 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.

Fig. 3: 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 and the eye.

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 two internal representations are mutually exclusive alternatives. Each of the representations can be convolved with the input for a modulation and selection of locations (or objects) within the image as target areas of visual selection. The input representation is the same as in standard bottom-up models (Itti, 2005). 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.

To start with the memory templates, the two-step model's visual memory contains representations of visual feature combinations (in a space of manifolds to also allow for edge representations) for objects and for scenes (or takes). If such a representation is retrieved (for the exact conditions under which retrieval takes place, see below), this memory representation can be used as a template to up-weigh repeated feature combinations as relevant during visual search within the image relative to the irrelevant non-repeated feature combinations. This conception of a retrieved search template is very similar to that of other top-down models of attention (cf. 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 (cf. Maxcey-Richard & Hollingworth, 2013). The two-step model thus uses gaze direction for segmentation (here: the extraction of objects in an image surround) 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. In this manner, the two-step model adapts to inter-individual variation of looking preferences and keeps track of them, without having to make additional assumptions.

As explained, if this is necessary the model assumes that human viewers can retrieve their stored matrices from visual memory and use a retrieved matrix as a search template for the direction of their attention. But when is this necessary? In the two-step model, we assume that whether or not one matrix will be concluded and stored and another matrix retrieved from visual memory depends on the temporal coherence of the optic flow across images. When technically classifying video material, two common characteristics are used: the spatial detail and temporal activity (ITU-T P.910). Spatial detail, also referred to as Spatial Information (SI) is a measure for the amount of detail visible in video. It is calculated by applying a filter over each frame. Temporal Activity (or Temporal Information, TI) is a similar measure that calculates the difference between two subsequent frames on a pixel-by-pixel basis. Calculating these measures for edited video material, shows that TI shows peaks whenever cuts occur in edited videos (see Figure 4).

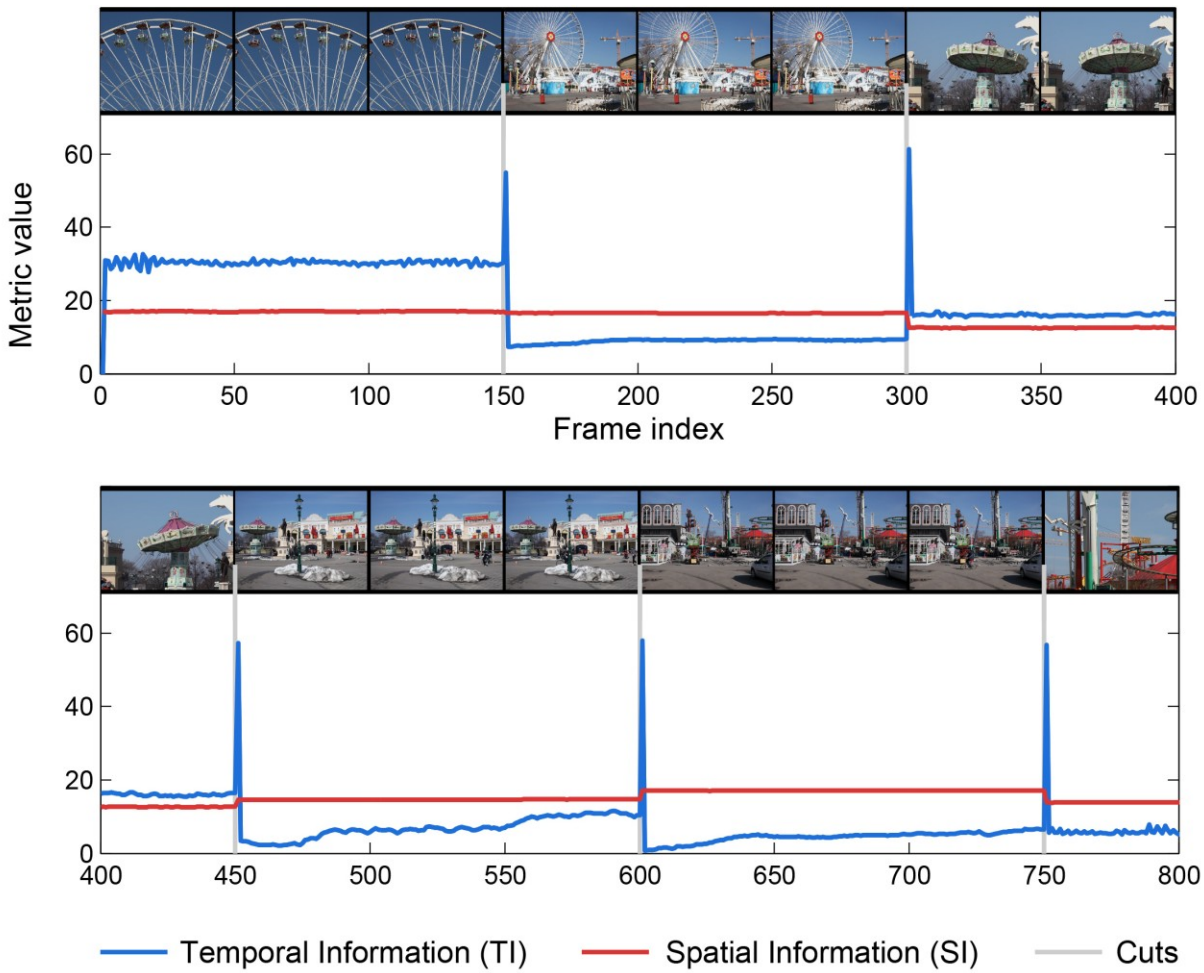

Fig. 4: Measures of Temporal Information (TI) and Spatial Information (SI) (cf. ITU-T P.910) for an example video clip featuring multiple takes. Cuts from one take to the next occurred every 150 frames (image stills depict the first frame in the clip and then every 50th frame). The result illustrates that TI (plotted in blue) shows peaks at the time of cuts. Instead, SI (as measure of spatial detail) does not change significantly across cuts. (The source video clip can be viewed from http://vimeo.com/76033466).

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. 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. This is achieved by calculating the differential (or first derivation) of the transformation rule as a function of 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 down-weighing 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.

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. To be precise, after a minimum of optic-flow coherence, one scene-specific or take-specific matrix will be retrieved from visual memory. We assume that, as a default, humans retrieve the most recent scene matrix (Maxcey-Richard & Hollingworth, 2013). This corresponds to the matrix of the take that the viewer has seen before a minimum of optic-flow coherence. Next, the retrieved scene matrix will be used as a top-down search template for the direction of attention in the image: After a minimum of optic-flow coherence, the two-step model thus predicts that the fixations should be mostly directed to repeated feature

combinations. Attention and hence the eyes should be attracted towards those areas of the visual image with a high resemblance to the retrieved feature matrix that currently serves as a search template. Attraction of attention (e.g., the gaze) should be proportional to the computed local similarity of the visual feature combinations in the input representation and the search templates. This happens with a slight delay, during a buffering fixation directed to a neutral image positions (e.g., its center; Dorr, Martinetz, Gegenfurtner, & Barth, 2010).

In this way, the two-step model capitalizes on feature priming as a second top-down principle for the direction of attention and uses it complementarily to surprise capture. Both of the two top-down principles take turns because they are the respective consequences of two mutually exclusive states of a joint steering variable: the temporal coherence function of optic flow. To note, it is assumed that in edited videos a high coherence of optic flow should correspond to the times between cuts (or within takes), whereas a minimum of the temporal coherence of optic flow reversely should correspond to the point in time of a cinematic cut (or the time between takes). This means that according to the two-step model, between cuts attention will be more likely directed to the novel visual features than to the repeated visual features because the novel features are informative. After a cut, however, re-orienting is required to make sense of the video stream's meaning. This in turn, leads to the attraction of attention towards the repeated visual features and objects, such as the visual landmarks, for the purpose of recognizing a scene as a continuation of the previous scene or as a novel scene.

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 model's free parameters (the content of the visual memory) will not be arbitrarily chosen but will be specified on the basis of empirical observation (i.e., will be measured as the feature values at fixated positions). The latter also implies that the two-step model is tied to empirical hypotheses, and can be tested in well controlled psychological experiments. In addition, the model can also be tested in more applied research areas, such as clinical diagnosis bases on visual motion (e.g., in ultrasound imaging), and QoE (quality of experience) assessment in entertainment videos.

3. EVIDENCE FOR THE TWO-STEP MODEL

3.1 The Top-Down Weighting of Coherent Optic Flow

The surprise-capture principle – that is, the attraction of the eyes by unexpected, newly added visual information as compared to expected and repeated visual information, outperforms the bottom-up model when predicting fixations within animated video games and movies (cf. Itti & 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 camera's (or the observer's) 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. Assuming that (1) optic flow is a typical consequence of camera motion or self-motion and (2) perception ultimately serves inferences about distal objects in the environment, optic flow is also a prime candidate for top-down suppression or filtering-out because in many instances optic flow would reflect an 'artifact' – that is, a property of the viewer's or camera's shifting viewpoint rather than a property of the distal visual stimulus in the environment.

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, Wolfe, & Klempen, 2001). Likewise, objects moving relative to the flow field pop out from the background (Rushton, Bradshaw, & Warren, 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, Nir, Lavy, Fuhrmann, & Malach, 2004) and human faces in particular (Foulsham, Cheng, Tracy, Henrich, & Kingstone, 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 & Christ, 2003; Becker & Horstmann, 2011).

3.2 Top-Down Directing of Attention by Templates Combining Features (Color and Shape)

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 (cf. 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 & Humphreys, 1989; Folk, Remington, & Johnston, 1992). 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 & Gelade, 1980). In this way human viewers could also search for landmarks that they have seen in the past to re-orient 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, Rasmussen, & Hollingworth, 2010; Chun & Jiang, 1998; Chun & Wagner, 2006). Similar advantages have been demonstrated in the context of visual recognition under more natural conditions, with static photographs of natural scenes (Foulsham & Kingstone, 2013; Maxcey-Richard & Hollingworth, 2013; Valuch, Becker, & Ansorge, 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 or they were from an at least equally salient non-fixated area of the learned images. 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.

3.3 Feature Priming as a Top-Down Principle of Attention – Linking Repeated Features Across Time Enables 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. In line with the assumed role of repetition priming on eye movements, Valuch et al. (2013) observed that participants preferentially looked at scene positions that repeated from learning to recognition. These authors used static images that the participants had to learn while the participants' eyes were tracked. Later, during recognition only a subset of the learned images was repeated and participants had to discriminate old images, including perspective shifts of the old images, from new and mirrored images. Critically, in the recognition block, the participants showed a clear preference for looking at the repeated image parts (see Figure 5). For example, the new perspectives were created to exactly repeat the learned image in only one half of the pictures (on the left or on the right), and the other half of the recognition images contained information that was hitherto not presented to the participants. In this situation, the participants showed a clear preference to look at the repeated image parts.

Fig. 5. Gaze preference for repeated image parts during recognition of scenes after view changes. Images that were presented during learning, and during recognition, respectively, were smaller cutouts from originally larger source images. In this example, the alternative view was created by shifting the crop frame to the left border of the original source image. In this way, the recognition image repeated exactly half of the learned image's parts (i.e., the right half of the recognition image is equal to the left half of the learned image). During recognition, viewers fixated significantly more often and longer on the repeated image parts. Data and stimuli are based on Valuch, Becker, and Ansorge (2013).

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, Levin, & Cutting, 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, Brunick, & Candan, 2012).

3.4 Attention to Repeated Versus Novel Features are Complementary Top-Down Principles of Visual Attention

In line with different strategies within takes than across cuts, 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 (H. X. Wang, Freeman, Hasson, Merriam, & Heeger, 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.

4. OPEN QUESTIONS

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 and whether it can provide a signal to detect cuts. To address this question, one would need to develop a robust algorithm that segregates optic flow fields into global and local areas of higher or lower coherence of the signal across time. This requires a maximally sensitive algorithm that discriminates coherent optic flow from other forms of visual motion. One way how this problem could be addressed is by extending flow-driven spatial smoothness in terms of spatio-temporal regularizers (Weickert & Schnörr, 2001a, b) and a decomposition (Abhau, Belhachmi & Scherzer, 2009; Meyer, 2001) of optical flow into bounded variation and an oscillating component. The decomposition and the time regularization could be used to improve the estimation of the optical flow and to extract more information from a dynamic sequence. Due to that it would be possible to study in which way the components of the optical flow model affect visual attention, for example, to determine the thresholds at which regular patterns of optic flow take a different effect on the attraction of attention as compared to other forms of visual motion. Likewise, because the boundary between optic flow and other forms of visual motion is uncertain and arbitrary, the predicted patterns of eye movements as indices of attention should help to learn more about this boundary.

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 (see Burnham, 2007, for a review). 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. Critically, laboratory research has already shown that top-down search for color can be compromised where the participants have to search for more than a single color (Grubert & Eimer, 2013) or where it is necessary to quickly discriminate one relevant from several irrelevant colors (Ansorge & Becker, 2013). Assuming that exactly these capacities are required when looking for target objects consisting of color spectra among distractor objects with partially overlapping spectra, it is not certain whether and how color is used in a top-down way in more naturalistic images.

In addition, a lot more questions than answers arise with regard to the storage and usage of different take-specific top-down templates. First of all, it seems that visual working memory is of severely limited capacity (Luck & Vogel, 1997) so that the question is how precise any take-specific memory might be. A related question concerns the retrieval of a fitting template from memory. Above, we have discussed only the simplest situation in which one scene either repeats or does not repeat across cuts. However, human perceivers are also able to recognize parallel cutting where between cuts, the movie jumps back and forth between two different scenes. In this situation humans can easily recognize the streams as belonging to either of two scenes, even if a part of the alternative scene has been inserted between two takes from the same scene. In this situation, participants must somehow be able to retrieve more than just the most recent template for their decisions.

4.1 Potential Applications of the Two-Step Model

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), and QoE (quality of experience) assessment 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. For example, according to our model, after an ultrasound device has lost contact with the body, for the viewer it should be helpful if the last registered ultrasound image before the contact loss would be convolved with the first image after the contact has been reestablished, for the highlighting of the overlapping visual information in consecutive but interrupted images. This highlighting should be

helpful for the reorientation of the viewer because per definition the repeatedly looked-at image regions before and after the interruption must be a subset of all repeated image content before and after the cut. Also, 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 bearing 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. In general, the loss of information from videos can be compensated by impainting or extrapolation of information from previous frames for the prediction and reconstruction of the lost images. This strategy, however, is not feasible for the first frame after a cut. In the case of a cut, the complete set of pixels has to be coded and is otherwise lost. Researchers therefore already proposed to reduce the amount of back-up data needed for reconstructing images after a cut by integrating an attention model (Chen et al., 2007, May). For this purpose, so far, researchers used the previously mentioned bottom-up model of attention (Itti, Koch & Niebur, 1998) but clearly our model suggests using looked-at input instead. Again applications of our model based on pervasive eye-tracking could lead the way to the solution of this applied problem. Clearly, however, whether our model improves the situation in these instances is an open question.

5. CONCLUSION

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. 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, A. Z., & Scherzer, O. (2009). On a decomposition model for optical flow. *Lecture Notes in Computer Science, Energy Minimization Methods in Computer Vision and Pattern Recognition, 5681*, 126-139.
Abrams, R. A., & Christ, S. E. (2003). Motion onset captures attention. *Psychological Science, 14*, 427-432.
Ansorge, U., & Becker, S. I. (2013). Contingent capture in cueing: The role of color search templates and cue-target color relations. *Psychological Research*, in press.
Ansorge, U., Priess, H.W., & Kerzel, D. (2013). Effects of relevant and irrelevant color singletons on inhibition of return and attentional capture. *Attention, Perception, & Psychophysics*. doi:10.3758/s13414-013-0521-2.
Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences, 11*, 280-289.
Baron-Cohen, S. (1997). How to build a baby that can read minds: Cognitive mechanisms in mind reading. In S. Baron-Cohen (Ed.), (pp. 207-239). Hove, UK: Psychology Press.
Becker, S. I., & Ansorge, U. (2013). Higher set-sizes in pop-out search displays do not eliminate priming or enhance target selection. *Vision Research*, *81*, 18-28.
Becker, S. I., & Horstmann, G. (2011). Novelty and salience in attentional capture by unannounced motion singletons. *Acta Psychologica, 136*, 290-299.
126-139.
Böhme, M., Dorr, M., Krause, C., Martinetz, T., & Barth, E. (2006). Eye movement predictions on natural videos. *Neurocomputing, 69*, 1996-2004.
Born, S., Ansorge, U., & Kerzel, D. (2012). Feature-based effects in the coupling between attention and saccades. *Journal of Vision, 12(11)*:27.
Born, S., Ansorge, U., & Kerzel, D. (2013). Predictability of spatial and non-spatial properties improves perception in the pre-saccadic interval. *Vision Research, 91,* 93-101.
Brooks, D. I., Rasmussen, I. P., & Hollingworth, A. (2010). The nesting of search contexts within natural scenes: Evidence from contextual cueing. *Journal of Experimental Psychology: Human Perception and Performance, 36*, 1406-1418.
Burnham, B. R. (2007). Displaywide visual features associated with a search display's appearance can mediate attentional capture. *Psychonomic Bulletin & Review*, *14*, 392-422.
Carmi, R., & Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research, 46*, 4333-4445.
Chen, Z., Qiu, G., Lu, Y., Zhu, L., Chen, Q., Gu, X., & Wang, C. (2007, May). Improving video coding at scene cuts using attention based adaptive bit allocation. In *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on* (pp. 3634-3638). IEEE.
Chun, M. M., & Jiang, Y. (1998). Contextual cuing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*, 28-71.
Chun, M. M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science*, *10*, 360-365.
Cutting, J. E., Brunick, K. L., & Candan, A. (2012). Perceiving event dynamics and parsing Hollywood films. *Journal of Experimental Psychology: Human Perception and Performance, 38*(6), 1476.
Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36*, 1827-1837.
Dorr, M., Martinetz, T., Gegenfurtner, K.R. & Barth, E. (2010) Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision, 10(10)*: 28.
Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*, 433-458.
Fletcher, P. C., Anderson, J. M., Shanks, D. R., Honey, R., Carpenter, T. A., Donovan, T., Papadakis, N., & Bullmore, E. T. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nature Neuroscience, 4*(10), 1043-1048

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 1030-1044.
Forster, S., & Lavie, N. (2009). Harnessing the wandering mind: the role of perceptual load. *Cognition*, *111*, 345–55.Foulsham, T., & Kingstone, A. (2013). Fixation-dependent memory for natural scenes: An experimental test of scanpath theory. *Journal of Experimental Psychology: General, 142*, 41-56.
Foulsham, T., Cheng, J. T., Tracy, J. L., Henrich, J., & Kingstone, A. (2010). Gaze allocation in a dynamic situation: Effects of social status and speaking. *Cognition, 117*, 319-331.
Frintrop, S., Rome, E., & Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP),* 7(1), 6.
Gibson, B. S., & Amelio, J. (2000). Inhibition of return and attentional control settings. *Perception & Psychophysics*, *62*, 496-504.
Grubert, A., & Eimer, M. (2013). Qualitative differences in the guidance of attention during single-colour and multiple-colour visual search: Behavioural and electrophysiological evidence. *Journal of Experimental Psychology: Human Perception and Performance*, in press.
Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2006). Intersubject synchronization of cortical activity during natural vision. *Science, 303*(5664), 1634-1640.
Hochberg and Broocks. (1996). The perception of motion pictures (revised). In M. P. Friedman & E. C. Carterette (Eds.), *Cognitive ecology* (pp. 205-292). San Diego, CA: Academic Press.
Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence, 17*, 185–203.
Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition, 12*, 1093-1123.
Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*, 1295-306.
Meyer, Y. (2001). Oscillating patterns in image processing and nonlinear evolution equations. *University Lecture Series. The fifteenth Dean Jacqueline B. Lewis memorial lectures, American Mathematical Society, 22*.
Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*, 1489–1506.
Itti, L., Koch, C., & Niebur E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions of Pattern Analysis and Machine Intelligence, 20*, 1254–1259.
Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences, 4*, 138-147.
Koenderink, J. J. (1986). Optic flow. *Vision Research, 26*, 161-179.
Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research, 35*, 1897-1916.
Lee, D. N., & Kalmus, H. (1980). The optic flow field – the foundation of vision. *Philosophical Transactions of the Royal Society B – Biological Sciences, 290(1083)*, 169-179.
Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*, 279–281.
Maxcey-Richard, R., & Hollingworth, A. (2013). The strategic retention of task-relevant objects in visual short term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 760-772.
Maljkovic, V., & Nakayama, K. (1994). Priming of pop-out: I. Role of features. *Memory & Cognition*, *22*, 657-672.
Meeter, M., & Olivers, C. N. L. (2005). Intertrial priming stemming from ambiguity: A new account of priming in visual search. *Visual Cognition*, *13*, 202-222.
Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2010). Clustering of gaze during scene viewing is predicted by motion. *Cognitive Computation, 3*, 5-24.
Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32*, 3-25.
Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. In H. Bouma & D. G. Bouwhuis (Eds.). *Attention and Performance Vol. X*, (pp. 531-556). Hillsdale, NJ: Erlbaum.
Priess, H.-W., Born, S., & Ansorge, U. (2012). Inhibition of return after color singletons. *Journal of Eye Movement Research*, *5*(5), 1-12.
Ranganath, C., & Rainer, G. (2003). Neural mechanisms for detecting and
remembering novel events. *Nature Reviews. Neuroscience, 4*(3), 193–202.
Royden, C. M., Wolfe, J. S., & Klempen, N. (2001). Visual search asymmetries in motion and optic flow fields. *Perception & Psychophysics, 63*, 436-444.
Rushton, S. K., Bradshaw, M. F., & Warren, P. A. (2006). The pop-out of scene relative object movement against retinal motion due to self-movement. *Cognition, 105*, 237-245.
Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences*, *15*, 319–326.
Smith, T. J., Levin, D. T. & Cutting, J. (2012). A Window on Reality: Perceiving Edited Moving Images. *Current Directions in Psychological Science, 21,* 101-106.
Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics*, *51*, 599-606.
Theeuwes, J. (2010). Top-down and bottom-up control of visual attention. *Acta Psychologica*, *123*, 77-99.
Torralba, A., Oliva, A., Castelhano, M. S., & 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*, 766–786.
Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97-136.
Valuch, C., Becker, S. I., & Ansorge, U. (2013). Priming of fixations during recognition of natural scenes. *Journal of Vision*, *13(3):3*, 1-22.
Walther, D., & Koch C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*, 1395–1407.
Wang, Z. G., & Klein, R. M. (2010). Searching for inhibition of return in visual search: A review. *Vision Research, 50*, 220-228.
Wang, H. X., Freeman, J., Merriam, E. P., Hasson, U., & Heeger, D. J. (2012). Temporal eye movement strategies during naturalistic viewing. *Journal of Vision, 12(1)*: 16.
Weickert, J., & Schnörr, C. (2001a). A theoretical framework for convex regularizers in PDE-based computation of image motion. *International Journal of Computer Vision, 45*, 245–264.
Weickert, J., & Schnörr, C. (2001b). Variational optic flow computation with a spatio-temporal smoothness constraint. *Journal of Mathematical Imaging Vision, 14*, 245–255.
Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, *1*, 202-238.
Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, *5*, 1-7.
Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review, 115,* 787-835.

Received *** 2013;