Directing Attention and Influencing Memory with Visual Saliency Modulation

Eduardo Veas[1], Erick Mendez[1], Steven Feiner[2], Dieter Schmalstieg[1]

[1]Institute for Computer Graphics and Vision
Graz University of Technology, Graz, Austria
{veas, mendez, schmalstieg}@icg.tugraz.at

[2]Department of Computer Science
Columbia University, New York, USA
feiner@cs.columbia.edu

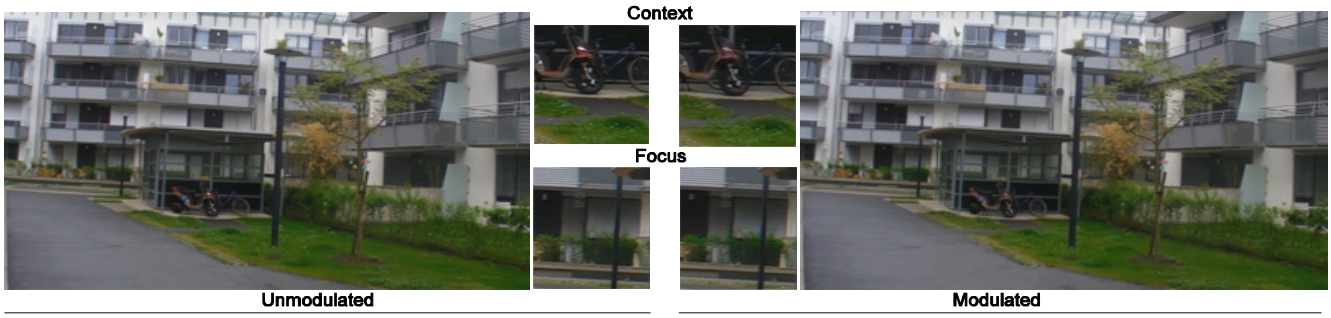

Figure 1. Comparison between modulated and unmodulated video. (Left) Frame and details from an unmodulated video. (Right) Same frame and details after modulation. Differences may be seen when compared side by side, but evidence of modification is difficult to see when viewing the modulated version in isolation.

ABSTRACT

In augmented reality, it is often necessary to draw the user's attention to particular objects in the real world without distracting her from her task. We explore the effectiveness of directing a user's attention by imperceptibly modifying existing features of a video. We present three user studies of the effects of applying a saliency modulation technique to video; evaluating modulation awareness, attention, and memory. Our results validate the saliency modulation technique as an alternative means to convey information to the user, *suggesting* attention shifts and *influencing* recall of selected regions without perceptible changes to visual input.

Author Keywords

Saliency, Visual Attention, Augmented Reality.

ACM Classification Keywords

H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Screen design (e.g., text, graphics, color)*; H.5.m [**Information Interfaces and Presentation**]: Miscellaneous; I.4.3 [**Image Processing and Computer Vision**]: Enhancement.

General Terms

Human Factors, Experimentation.

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

INTRODUCTION

Augmented reality (AR) applications intended to call attention to real objects often do so by overlaying on the real world highlighting effects or virtual objects such as arrows. At times, it would be desirable that these effects were more subtle, in part to avoid exacerbating perceptual issues inherent to AR (such as depth perception, occlusion and visual clutter), but mostly in cases where the objects highlighted are secondary to the user's task. In many cases, the application needs to appeal to post-perceptual processes, to tell the user that a particular object is somehow related to their current task, but without alerting or interrupting the user's workflow. For example, an environmental scientist visualizes simulation results overlaid on a mountain landscape, and the application wants to highlight the sensors that contributed data to the simulation. Similarly, a user watching a remote video feed from a multi-camera system needs to be reminded of the locations of the viewing camera, of other cameras, and of interesting objects [19]. Or, a panoramic 3D system for navigation may need to draw a tourist's attention towards buildings or landmarks along a path.

The technique presented in this paper offers an alternative means to convey information to the user. We investigate its effectiveness to 1) direct attention to a selected region of the visual input and 2) influence the recall rate of certain objects, 3) without the user becoming aware of any modifications. The literature on psychology and vision identifies saliency as a model of attention [8]. Moreover, attention influences memory at different stages of processing [1]. Thus, we assume that by manipulating the saliency of a region in the visual input, we can potentially influence at-

tention and memory. We apply a saliency modulation technique (SMT) to modify videos so that a region of our selection contains the highest saliency. The SMT enables an AR approach known as "mediated reality," in which existing features of the real-world image are modified, instead of adding discrete new objects. Our primary contribution is to show, by modulating prerecorded video in a lab setting, the potential for developing AR user interfaces that imperceptibly direct a user's attention toward other parts of their environment that are auxiliary to the user's task.

We performed three studies, measuring modulation awareness, attention, and memory. The *modulation awareness* study finds the largest amount of modulation we can apply that is imperceptible to the viewer. The *attention* study evaluates whether this modulation threshold shifts attention towards selected regions of videos. The *memory* study evaluates whether it increases recall for selected objects. Our results indicate that regions modulated with the SMT will draw a first fixation faster than without modulation. Moreover, modulation can increase recall for selected objects. In summary, the SMT can significantly shift attention to selected areas and influence memory of selected objects from a video in a way that is imperceptible to the viewer.

RELATED WORK

Visual salience (or *visual saliency*) is the distinct subjective perceptual quality that makes some items in the world stand out from their neighbors and immediately grab our attention [8]. It refers to the evolved process in primates and other animals that restricts complex object recognition to small areas or objects at any one time that are analyzed serially. Saliency is commonly agreed to have bottom-up and top-down factors. *Bottom-up*, (memory-free, stimulus-based) factors refer to pure sensory information, such as a bright light suddenly appearing in front of us. *Top-down* (memory-bound, goal-based) factors involve a conscious effort, such as ignoring more salient stimuli while carefully scanning a book index. This paper focuses on *bottom-up* factors, which announce to the organism whether a location differs enough from its surroundings to warrant attention.

Measurements of the attention process of an organism are typically focused on stimulus-only factors. The most influential work on understanding this was done by Treismann and Gelade [18], and by Koch and Ullman [11]. Koch and Ullman, in particular, proposed the idea of a single map that is a combination of individual salient contributions; the normalized result is referred to as the *saliency map*. They state that the saliency at a given location is determined primarily by how different this location is from its surround in properties such as color, orientation, motion, and depth.

Saliency and Visual Attention

There is much evidence indicating a correlation between visual attention and the saliency map. Ouerhani et al. [14] and Santella et al. [16] used an eye tracker to confirm the relationship between the saliency map and human visual attention. Lee et al. [12] went one step further by using the saliency map to track objects being attended to by the user.

Practically any change made to an image will modify its saliency map. Blurring, (de)saturating, harmonizing, and distorting are operations that implicitly change the saliency of an image. Recent research has focused on directing attention through saliency manipulation for volume rendering [9], non-photorealistic stylization [16], and geometry [10]. These works concentrate on creating salient features; in contrast, our work receives an *existing* image as input and outputs a modified image whose saliency is manipulated without adding new features.

Closest to our intentions is the work by Su et al. [17] on de-emphasizing distracting image regions and by Bailey et al. [2] on subtle gaze direction. Su et al. focused on so-called second-order saliency by modulating variations in texture to redirect the user's attention to specific locations. Bailey et al. apply first-order modulations to the focus, only when the user is not looking there, as determined by an eye tracker. In contrast, our technique works with dynamic live video and can thus support augmented reality applications with arbitrary scenes and without requiring an eye tracker.

Saliency and Memory

There is a two-way relation between attention and memory that has been widely studied in the past [1][7][4]. Awh et al. [1] identified experiments leading to the conclusion that attention influences processing during both early sensory and post-perceptual stages. They also collected evidence supporting that the same attentional processes that facilitate early sensory identification of new information are recruited for active maintenance of information in memory. Two recent studies have proven the influence of saliency in memory, albeit with different results regarding the reasons. Berg and Itti [3] concluded that salience contributes to memory by influencing overt attention. They had participants examine a shopping-related scene for 2s and then asked if a target item was contained in the scene. They found that fixation times, but not saliency, influenced performance. Fine and Minnery [5] found that the influence of saliency extends beyond oculomotor behavior to higher order centers involved in spatial working memory. They presented participants with maps that included a number of icons to memorize. After a pause, participants had to drag each icon to its original position. They found that participants attended to icons equally regardless of their saliency (quantified using the model from Itti et al. [6]), but errors in placement were significantly reduced for salient icons. Thus, results could not be explained by a biasing of overt attention. Both cases support the fact that saliency influences memory. We assume that by actively modifying an object's saliency, we can influence memory.

SALIENCY MODULATION TECHNIQUE

We apply a recently developed SMT capable of manipulating the saliency of a video [13]. The SMT works at interactive framerates. For each frame, the SMT computes a saliency measure on every fragment according to a hierarchical multi-channel contrast measure [6]. It then modifies the image, changing contrast in lightness and color to have

the highest attention salience inside a designated focus region. Changes are applied so that spatial and temporal coherence are respected.

In detail, the SMT works by analyzing and modulating conspicuities in three dimensions: lightness (L), red–green color opponents (O_r), and blue–yellow color opponents (O_b). Each frame is first converted to CIE L*a*b space, thereby obtaining the values for each dimension $k \in \{L, O_r, O_b\}$. A pyramid of images is created with p levels. Modulation progresses from coarse levels to fine levels of the image pyramid. This allows changes affecting a large region to occur early in the process, while later steps progressively refine the result, thus introducing less noticeable artifacts. For each level, *analysis* and *modulation* steps are carried out iteratively for each dimension k.

Saliency analysis. During this step, the conspicuities of the image are computed to measure the naturally salient objects in the scene. A conspicuity is given as a signed sum of the center–surround differences at multiple scales of an image pyramid. The conspicuity c_k is the defined as:

$$c_k = \frac{\sum_{n=0}^{n=2} \sum_{m=n+3}^{m=n+4} k_n - k_{n+m}}{p},$$

where $p = 6$, and k_i is the conspicuity $k \in \{L, O_r, O_b\}$ at mipmap level i.

The conspicuity c_k is normalized using the global conspicuity maxima [12]. The normalized conspicuity $\hat{c}_k$ is:

$$\hat{c}_k = \frac{c_k}{\max(c_k)},$$

where $k \in \{L, O_r, O_b\}$.

Saliency modulation. Given a dimension $k \in \{L, O_r, O_b\}$, let $\hat{c}_k$ be the normalized conspicuity of a location and t_k be the threshold of the conspicuity, a floating point number that governs the amount of modulation. A modulation adjustment m_k is calculated for this location as,

$$m_k = \begin{cases} 0 & \hat{c}_k < t_k \\ \hat{c}_k - t_k & \text{otherwise} \end{cases}.$$

For a feature value f_k of a location, the modulated value f'_k is calculated by applying the modulation m_k in order to increase the conspicuity of the focus, and correspondingly decrease that of the context. Thus,

$$f'_k = \begin{cases} f_k + m_k & \text{if the location is marked as focus} \\ f_k - m_k & \text{otherwise} \end{cases}.$$

Modulation is performed in the order of sensitivity of the human visual system [20]: first, lightness is modulated, then red–green opponents, then blue–yellow opponents. Note that other contributors to saliency remain unaffected (e.g., motion, size, and orientation). Finally, the image is converted from CIE L*a*b to RGB. See Mendez et al. [13] for implementation details. Our contribution is in applying thresholds so that modulation is imperceptible.

METHODOLOGY

To prepare the stimuli for the awareness and attention studies, we recorded ~10h of video under various situations (indoors, outdoors, night, day, with moving objects, free moving camera, panning camera). The idea was to have a manageable variety of videos that represented day-to-day situations. From these, we extracted clips, each lasting ~10s, with the restriction that no human body parts appear in the clips because they represent a high attention sink. Videos were recorded at a resolution of 1280×720 at 29.97fps and presented without resizing and uncompressed to avoid interpolation artifacts by the graphics card. For each experiment, we recruited a balanced number of participants from the university population and the general public. All participants had normal or corrected to normal vision, and were screened for color-sensitivity deficiencies by an on-screen Ishihara test. We used an SMI desktop-mounted eye tracker, operating at 60 Hz. Stimuli were presented on a 19″ monitor at 70cm from the participant. A chin rest was used to limit head movements. All studies were performed in an empty office with lights off, and windows and doors closed, to minimize attention distracters.

Focus regions (FR) for the awareness and attention studies were chosen by analyzing the videos and selecting low salience regions. The selection methodology is presented in the next section. Each clip contained one or more FR. Each FR was visible for at least 2s.

EXPLORATORY STUDY: MODULATION AWARENESS

In the SMT presented above, the *amount* of modulation is governed by a threshold (t_k) for each modulation dimension. Thus, the SMT can be configured to produce different modulation thresholds (see Figure 2). Our initial concern was how to apply the SMT so that the viewer is unaware of the manipulation. In other words, we were seeking the *maximum* modulation that is imperceptible to the viewer. To investigate viewers' attitudes towards modulation, we conducted a series of studies on modulation awareness.

A threshold is a floating point value in the [0...1] range. To reduce the search space, we discretized this range into a set of seven samples. Additionally, we used the extreme values 0 (no modulation) and 1 (full modulation), for a total of nine thresholds. We performed three studies to investigate the appropriate modulation threshold. A challenge in these studies is that participants need to evaluate different modulation thresholds for videos by actively checking for visual manipulations in them, a goal-based task. This type of task is known to modify the gaze path of participants and suppress stimulus-based attention. Thus, analysis of attention cannot be performed at the same time as the study on awareness of modulation.

First Pilot Study

Our intention for the first pilot was to identify and discard thresholds for which modulation is clearly perceivable, thus reducing the search space for a subsequent formal study. The stimulus was a series of 18 clips, two for each of the nine modulation thresholds, lasting ~10s each.

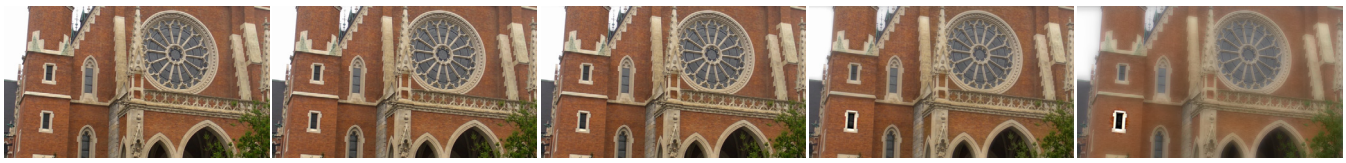

Figure 2. Modulation thresholds. Frame from a video modulated to emphasize a window using different modulation thresholds. From left to right, modulation thresholds zero (no modulation), three, four, five, eight (full modulation).

Three people (all male, ages 28, 33, and 35) participated in this study. Participants were requested to look at the videos and verbally rate each of them on a 7-point Likert scale for naturalness (where 1=natural corresponded to *the video is as it came from the camera*, and 7=unnatural meant *the video has been manipulated to such an extent that it feels unnatural*). The videos were shown in randomized order, in two sets of nine with a short break in-between. It is important to note that participants had to judge each video in isolation and the videos for each modulation threshold were different. Therefore, participants were not given the chance to compare a modulated video with the original version.

Analysis and Results

We did not perform any statistical analysis in this set due to its small sample size. We confirmed, however, that the higher the modulation threshold; the higher the score given by the participants (see Figure 3 top). Thresholds 0–5 scored below the middle of the scale (*somewhat unnatural*). In fact, thresholds 0 and 4 had an average score of 3.

Second Pilot Study

In the first pilot study, participants judged each modulation threshold in isolation. This raised the doubt of whether they would detect a difference if they were given the chance to see both modulated and unmodulated versions of the same stimulus. The goal of this pilot was to verify whether participants could notice a difference between modulated and unmodulated images. We randomly selected screenshots from the stimulus videos. These were presented in pairs with a change-blindness break in between, following the setup suggested by Rensink et al. [15]. For each pair, the images were presented in the order FBFBSBSB, where F corresponds to first image shown for 240ms, B to blank image shown for 320ms, and S to second image shown for 240ms. There were nine change-blindness sets, one for each threshold considered. We modulated two images for each of the nine modulation thresholds, totaling 18 image pairs.

Three participants took part in this study (2 male, 1 female, ages 28, 29, and 24). They were instructed to observe the images and state whether or not the images were different. Each participant saw each of the 18 image pairs once. The presentation of the image pairs was randomized. As suggested by Rensink et al., each change-blindness pair was presented for 60s. Participants, however, had the possibility to interrupt the sequence by stating a judgment.

Analysis and results

We did not perform statistical analysis for this set due to its small sample size. Figure 3 (bottom) shows responses for this study as stacked bars. We interpreted each affirmative response as a value of 1, and each negative response as a value of 0. As shown in the figure, the pair zero–zero was always correctly judged as being unmodulated (it never received an affirmative response). Pairs zero–seven and zero–eight were also always correctly judged as being modulated (6 affirmative responses each). Intriguingly, pair zero–four was graded higher than zero–five and zero–six.

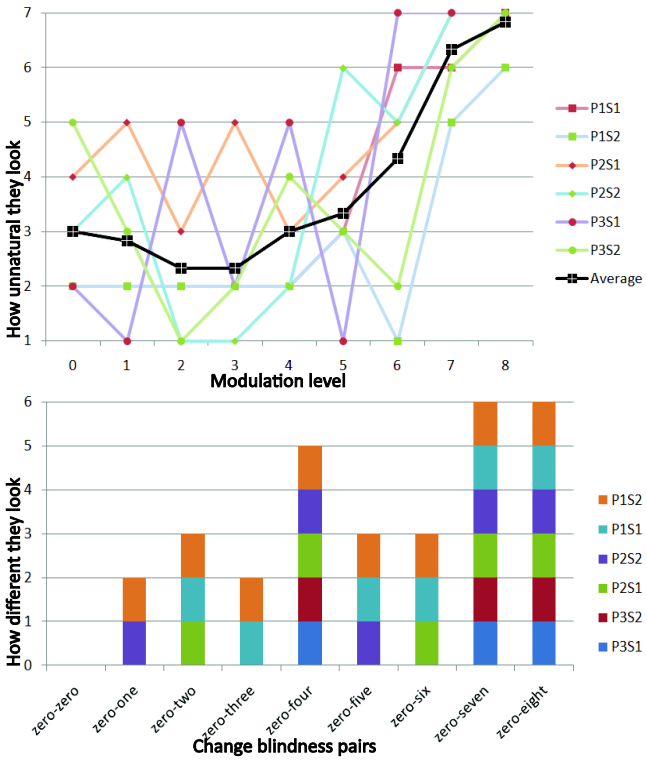

Figure 3. Responses of both pilots. (Top) Responses corresponding to the first pilot study. (Bottom) Responses corresponding to the second pilot study. Notation *PmSn* means *Participant m, set n*. Notation *zero–number* means a pair with the same image unmodulated and modulated at threshold *number*.

Formal awareness study

Building on the pilot studies, we conducted a formal study to further evaluate the reaction of people to modulation thresholds 3–5. Our aim was to verify that the threshold used in later experiments was imperceptible.

Method

The stimuli for the awareness study were obtained using the same 20 clips used in the attention experiment presented below. Therefore, three candidate thresholds (3, 4, and 5) plus the control (no modulation) times the 20 stimulus videos resulted in 80 video–threshold pairs. We arranged the videos so that each video–threshold pair was seen by four

participants. Each participant watched each video with a randomized modulation threshold. No participant watched the same video twice with different modulation thresholds.

We recruited 16 participants for this study (12 male, 4 female, 18–35 years old, $\overline{x}$=27.8), none of whom participated in the subsequent experiments. The procedure and instructions were the same as those described for the first pilot.

Analysis and Results

To analyze results, we considered the four modulation thresholds (0, 3, 4, and 5) as related samples. We then conducted three Wilcoxon signed tests for two related samples, to determine whether participants noticed significant damage to the videos compared to the ground truth. Our pair samples were zero–three, zero–four, and zero–five. We applied a Bonferroni correction to account for the number of pair samples and keep the α level below .05. The analysis showed no significant difference for any of the pairs. Thus, there was no evidence that the general population would be able to distinguish which videos had been modulated and which had not. However, we decided to take a somewhat conservative approach and use threshold four for our modulation procedure. Figure 1 illustrates the results of modulation. The left image was obtained from the unmodulated video (first condition). The right image was obtained from the video modulated with threshold 4 (second condition). When comparing both images side-by-side, changes are barely perceptible. If, however, one is allowed to see only the modulated video in isolation, the changes become imperceptible. The insets in Figure 1 show a detailed comparison of the changes. Observe that the focus after modulation has slightly more vivid colors and more contrast, while the context has slightly duller colors and less contrast.

ATTENTION EXPERIMENT

The goal of this study was to verify, through use of an eye tracker, whether the SMT can direct the visual attention of participants to selected regions. Here, we assume that a participant's visual attention can be characterized by their eye gaze. As stimuli for this experiment, we selected 20 clips lasting roughly 10s each.

Hypotheses

H1: The time before the first fixation on the FRs will be smaller for the videos modulated with our procedure than for the original unmodulated videos.

H2: The fixation time in the FRs (i.e., sum of durations of all fixations on the focus) will be higher for the videos modulated with our procedure than for the original unmodulated videos.

H3: The percentage of participants that have at least one fixation on the FR will be higher for the videos modulated with our procedure than for the original unmodulated videos.

Method

Since we wanted to compare eye-gaze for regions between unmodulated and modulated versions of a video, we used a between-subjects, repeated measures design with independent variable *modulation* (unmodulated, modulated), and dependent variables *time before first fixation*, *fixation time* (i.e., sum of durations for all fixations on the focus), and *percentage of participants with at least one fixation*.

We recruited 40 participants to take part in this experiment. They were divided into two conditions for the between-subjects setup (20 in the unmodulated condition, 20 in the modulated condition): (20 regions × 20 participants) = 400 trials per condition = 800 trials total.

Twenty participants (14 male, 6 female, 24–52 years old, $\overline{x}$ =31.4) took part in the first (unmodulated) condition. Each participant was provided with the following instructions:

You will sit in front of a computer screen. We will display a series of short video clips. All you have to do is look at the clips. That's it!

This test is divided into two parts so you can have a break in between. Your eye gaze will be tracked with a non-wearable system. It will be using an infrared camera and light placed in front of you. Infrared light is invisible to the eye and poses no harm to you.

Care was taken not to mention the number of video clips in order to avoid counting (which would trigger a top-down task). It was emphasized that there was no task and that all that was required was to watch the clips. The eye tracker was calibrated for each participant before the stimuli were presented. Each participant watched each of the 20 unmodulated videos once, in random order. Between videos, a blank slide was shown for 2000ms.

By analyzing eye-gaze data from the first condition, we determined a visually unattended region for each clip in the unmodulated stimuli. We define *unattended regions* as those that have fewer than five fixations by less than twenty percent of the participants. These unattended regions were then designated as the FRs of the study. To increase the saliency of FRs for the second condition, the clips were modulated with the SMT at threshold 4, as suggested by the awareness study. The clips derived through this process were used as stimuli for the second condition.

Twenty participants (14 male, 6 female, 25–42 years old, $\overline{x}$ =32.1) took part in the second (modulated) condition. They went through the same procedure as in the first condition, the only difference being that the stimuli were modulated.

Analysis

Analysis was performed with independent samples *t*-tests whenever our data satisfied the condition of normality and with Mann–Whitney *U* tests otherwise. In both cases, tests were one-tailed. Two-tailed tests would be able to indicate whether there is a significant difference between both conditions, but not whether this difference is in the intended direction (increasing attention). A Shapiro–Wilk test indicated that the data for H1 and H3 satisfied normality. However, the data for H2 did not. We adjusted α levels using a Bonferroni correction to ensure a level of .05.

Results for H1. The results of the one-tailed *t*-tests indicate that the mean values of the second condition (modulated) were significantly smaller than the mean values of the first condition (unmodulated), t (35) = 2.916, p < .01. Therefore, the mean duration before the first fixation on the FRs for participants in the second condition (M = 9231.58, SD = 468.16) was significantly smaller than that of the participants in the first condition (M = 9638.86, SD = 363.10).

Results for H2. There is no significant difference in total fixation time between the unmodulated (M=26.79) and the modulated (M=43.31) conditions (Mann–Whitney *U*=99.0, n1=17, n2=20, p=.3, one-tailed). Despite the lack of normality of the data for this hypothesis, we performed a *t*-test confirming the results t (35) = –2.117, p = .02. Therefore, the mean total fixation time for participants in the second condition (M = 43.31, SD = 24.32) was not significantly different from that of the participants in the first condition (M = 26.79, SD = 22.82).

Results for H3. The results of the one-tailed *t*-tests indicate that there is no significant difference in the number of participants that had at least one fixation between the unmodulated and the modulated conditions, t (35) = –2.028, p = .05. Therefore, the mean number of participants with at least one fixation in the second condition (M = .13, SD = .08) was not significantly higher than that of the participants in the first condition (M = .08, SD = .06).

As can be seen, H1 proved statistically significant; however, we were unable to find significant differences for H2 and H3. We further examined the gaze data of our participants to try to find consistent failures in our modulation procedure. By visually analyzing heat maps of our videos in the second condition, we found what seemed to be a consistent pattern where our modulation procedure failed: whenever the camera panned directly away from the FR, the technique seemed to be unable to attract fixations. This did not happen on videos where the camera was static, or whenever the panning was not directly away from the FR. Subsequently, we filtered out the information from FRs that fit this criterion (5 regions out of 29 were excluded). Then we repeated the analysis. Once again, we performed a Shapiro–Wilk test to verify the normality of our filtered data. The results indicated that the filtered data for H1 and H3 satisfied normality, but the filtered data for H2 did not.

On the filtered data for H1, the results of the one-tailed *t*-tests indicate that the mean values of the second condition (modulated) were significantly smaller than the mean values of the first condition (unmodulated), t (35) = 3.386, p < .01. Hence, the mean duration before the first fixation on the FRs for participants in the second condition (M = 9126.98, SD = 499.44) was significantly smaller than that of the participants in the first condition (M = 9616.66, SD = 352.59). On the filtered data of H2, results showed a significant difference in the total fixation time between unmodulated (Mean=14.03) and modulated (Mean=23.23) conditions (Mann–Whitney *U*=85.5, n1=17, n2=20, p< .01, one-tailed). Despite the lack of normality of the data for this hypothesis, we performed a *t*-test, which confirmed the significant difference in total fixation time between conditions t (35) = –2.659, p < .01. Consequently, the mean total fixation time for participants in the second condition (M = 49.52, SD = 26.04) was significantly higher than that of the participants in the first condition (M = 27.65, SD = 23.55). On filtered data for H3, the results of the one-tailed *t*-tests indicate that the mean values of the second condition were significantly higher than the mean values of the first condition, t (35) = –2.478, p < .01. Consequently, the mean number of participants with at least one fixation in the second condition (M = .15, SD = .09) was significantly higher than that in the first condition (M = .08, SD = .06).

Discussion of Attention Experiment

As can be seen from the analysis, we could always draw the eye gaze of participants significantly sooner with our modulation technique. However, once we filtered out situations in which the camera panned directly away from the FR, analysis revealed additional effects of the SMT in the modulated condition. On filtered data, the average duration before the first fixation on the FRs was significantly shorter. We could also retain the visual attention of participants for a significantly longer time. And finally, the number of participants with at least one fixation on the FRs of the modulated videos was significantly larger than for the unmodulated videos. It is difficult to illustrate the accumulated fixations on a region in a video, since the fixations on a region are spread out throughout the duration of the clip. Nevertheless, Figure 4 illustrates one frame of one video in both conditions, showing eye fixations accumulated over multiple frames, in which the effects of the SMT are clear. The image on the left comes from the unmodulated video and the image on the right from the modulated video. A white outline denotes the position of the FR. In the general case, however, the effect is not this apparent throughout the entire duration of the video.

Figure 4. Heatmaps of the user studies. (Left) Unmodulated condition. (Right) Modulated condition. This is a handpicked example chosen to illustrate the effect of the SMT.

The technique was not always effective for each of the video clips, nor for each of the participants in the tests. In the general case, the SMT will draw a first fixation faster than without modulation. In cases where the camera is not moving away from the focus regions, the number of participants that had at least one fixation in the focus region also increased, and the fixation time was significantly higher. Thus, we can state that attention direction with SMT was successful.

MEMORY EXPERIMENT

The goal of this experiment was to assess whether the SMT increases recall of selected objects in the video without suppressing recall for others. With the aim of comparing recall for regions between unmodulated and modulated videos, we used a between-subjects, repeated measures design with independent variable *modulation* (unmodulated, modulated), and dependent variable *recall hits*.

In order to prepare the stimuli, we recorded ~ 2h video in a furniture store and extracted two clips (identified as video *A* and video *B*) lasting 1m each. These clips include people walking by, but no faces. The choice of location ensured the appearance of many different objects in the videos.

Hypotheses

H4: There is no significant difference in recall hits between the first condition (unmodulated) and the second condition (modulated) for recalled objects.

H5: There is a significant difference in recall hits between the first condition (unmodulated) and the second condition (modulated) for non-recalled objects.

Hypothesis H4 concerns losses caused by the technique in the normal condition in terms of suppressing recall of normally recalled regions. H5 concerns gains due to the technique in terms of increasing recall of selected regions.

Prerequisites: Recall Study

The memory experiment requires a set of regions that appear in each video from which participants would select those they remember. These regions are associated with objects and are regarded as objects for the rest of the discussion. We expected to be able to determine the set of objects by examining the videos using Itti's model for saliency, but the results were mostly coarse, and would not help identify individual objects. We then decided to use a mixed approach in which we preselected some regions based on visual inspection and validated them by means of a pilot study. Thus, we visually examined the videos and selected scenes containing both low and high salience objects. Factors for scene selection included being clearly visible for an acceptable amount of time (about 2s), and that the objects in it be clearly distinguishable. We selected 18 scenes in total and, for each we extracted one object with high saliency and one with low saliency. Pictures of these objects were printed on 36 cards, each 11cm × 10cm.

To refine the set of objects, we carried out an exploratory study with six participants (5 male, 1 female, ages 25–35), who did not participate in any subsequent test. The procedure and apparatus were the same as those for the formal memory experiment. Based on eye-gaze analysis and on recall hits, seven scenes were removed, and three objects were changed in the remaining scenes resulting in deck *A* with 10 cards from video *A,* and deck *B* with 12 cards from video *B*. We classified five objects from deck *A* and seven from deck *B* as highly salient. The remaining objects were classified as having low saliency. This classification served as a control, as the experiment assumes that objects with high saliency will have high recall hits. We added five and six distracter objects to decks *A* and *B*, respectively, to assess whether a participant was picking cards randomly.

Method

The same 40 individuals that participated in the attention experiment took part in the memory experiment. Participants were divided into two conditions for the between-subjects setup (20 unmodulated, 20 modulated): (22 objects × 20 participants) = 440 trials per cond. = 880 trials total.

For each condition, videos *A* and *B* were shown in interleaved order; so that 10 participants experienced video *A* first and 10 participants experienced video *B* first. Before starting the experiment, participants were instructed to:

Observe the video and try to memorize the objects that you see. At the end you'll be presented with a deck of cards picturing objects printed from the video and you'll be asked to select those that you remember. Be careful, the deck of cards also contains objects that did not appear in the video.

Participants experienced the first video, and were subsequently presented with the corresponding deck of cards from which they could pick those objects that they remembered. A recall hit was recorded for an object if it was selected by a participant. After a short break, the same procedure was applied for the second video. For each video, participants answered a questionnaire in a 7-point Likert-scale format to assess the difficulty of the task. After finishing the procedure for the two videos, they answered general questions about the naturalness of the videos.

Based on analysis of the first condition, we classified objects as high recall (HR, recall higher than 60%) or low recall (LR, recall lower than 40%). The 40% and 60% thresholds were arbitrarily selected based on results of the first condition. Visual inspection of recall hits for this condition showed a gap in results: no object scored between 40% and 60%. Decks *A* and *B* had four HR objects each, totaling eight HR objects, all of which had been classified as highly salient. Three objects that had been classified as highly salient had low recall in the first condition, whereas objects classified as having low saliency all had low recall hits. In preparation for the second condition, videos *A* and *B* were modulated using the SMT to increase the salience of objects in LR. For the second condition, the only difference was the modulated stimulus; the procedure was the same.

Analysis

A Shapiro–Wilk test proved that the data for recall did not satisfy the condition of normality; the data are binary and not interval-scaled. Analysis was performed with Mann–Whitney *U* tests, due to their robustness under these conditions. Since our hypotheses focus on one side of the distribution, all the tests are one-tailed. We adjusted α levels with a Bonferroni correction to ensure a level of .05.

Results for H4. For objects $o \in HR$, mean recall hits in unmodulated (M=.69) and modulated (M=.68) conditions

show no statistical difference (Mann–Whitney U=1.1272e^4, n1=n2=160, p=.46, one-tailed). The result supports H4.

Results for H5. For objects $o \in LR$, the mean recall hits for unmodulated (M=.19) and modulated (M=.22) conditions show no statistical difference (Mann–Whitney U=3.78e^4, n1=n2=280, p=.15, one-tailed). There is not enough evidence to support H5.

To further analyze these results, we classified LR objects into those that increased in recall hits in the second condition, and those that did not show any change or showed a decrease in recall. We first confirmed the relationship between recall and attention, correlating recall with fixation count r(878) =.35, p<.001; and with fixation time r(878) =.666, p<.001. Then, we analyzed features of these objects that contribute to saliency and how they affect recall. We found moderate correlations between recall and size (in pixels) r(878) =.449, p=.032, and the average size in time of the region (% coverage × % visible time) r(878) =.428, p=0.042. We observed that objects $o \in LR$ that decreased in recall in the second condition were either < 2e^{4}px or appeared for less than 2s. Since the SMT cannot control contributions to saliency due to size or spatial frequency, we filtered data in LR based on these criteria, yielding two datasets: $LR' = o \in LR, size(o) > 2e^4 px$ and $MR = LR - LR'$. Subsequently, we analyzed the filtered data.

Object	o1	o2	o3	o4	o5	o6	o7
Unmodulated	2	2	3	2	8	1	8
Modulated	6	6	5	3	13	2	11

Table 1. Recall differences for objects in LR'

On filtered data for H5, objects $o \in LR'$, the mean recall for unmodulated (M=0.19) and modulated (M=0.31) conditions differed significantly (Mann–Whitney U=1.128e^4, n1=n2=160, p=.008, one-tailed). The result supports H5, meaning that objects that had low recall hits in the first condition significantly increased in score when modulated with the SMT (see Table 1). Furthermore, objects $o \in MR$ did not suffer a significant reduction in recall from unmodulated (M=.18) to modulated (M=.11) (Mann–Whitney U=6.660e^3, n1=n2=80, p=.05, one-tailed).

The results support H4 in the general case. This means that the SMT does not suppress recall for objects with otherwise high recall. The results did not provide enough evidence to support H5 in the general case. Nevertheless, for objects that cover more than 2e^{4}px and come into view for over 2s, the results showed a significant increase in recall. This suggests that the SMT increases recall of regions > 2e^{4}px with durations > 2s, without a significant loss to other regions.

Discussion of Memory Experiment

Exit interviews showed no difference in mean difficulty of the task between the first (M=5.28, STD=.987) and second conditions (M=4.98, STD=1.250), t(40)=1.190 p=.237 (2-tailed *t*-test). Furthermore, there was no difference in mean difficulty between video *A* (M=5.25, STD=1.171) and *B* (M=5.00, STD=1.086), t(40)=.990 p=.325 (2-tailed *t*-test).

The main contribution of this work is to show that the SMT introduces imperceptible changes to a video that increase recall of selected objects, without significantly reducing recall of others. The resource addressed, namely memory, is limited. In this study, participants in both conditions tended to remember the same number of objects, χ^2 (1, N = 880) = .39, p = .53. There is a tradeoff where the recall of some objects is reduced, while that of others is increased. In practice, our observations showed that a participant would recall on average five objects (at most eight) with certainty.

Figure 5. Scene extracted from the modulated condition. The high salience object (top left) had equally high recall hits (12) in both conditions. The low salience object (bottom left) had a score of 8 (LR) in the unmodulated condition and achieved a score of 13 (HR) in the modulated condition.

Some objects in HR decreased in recall, but not significantly (H4). Conversely, objects that were filtered out (objects $o \in MR$) also decreased in recall, albeit not significantly. In comparison, recall of objects $o \in LR'$ significantly increased. This comparison is between scores for the same object; it does not mean that we can increase recall hits of an object over those of another object. In particular, it does not mean we can increase recall hits of an inconspicuous object over those of a conspicuous object. The results merely show that the SMT increases the chances of an object being remembered. Having clarified this, there were two cases where the scores of an object in LR' increased to equal those of its scene counterpart in HR (Figure 5 shows one example). In both cases, recall hits in the first condition for the LR objects were at the 40% limit.

GENERAL DISCUSSION

The results presented in this paper indicate that the SMT can significantly shift attention to selected areas of a video, and it can increase recall of selected objects, without the viewer becoming aware of any manipulation. This provides strong evidence that the technique can influence the viewer's experience of a video at different levels of processing: it has applications in stimulus-based conditions (bottom-up) and task-based conditions (top-down).

Experiment Design

Since our main motivating field is AR, the studies were performed in videos, rather than still images. Using still images would not prove that the SMT works in videos,

which contain motion, a contributor to saliency that the SMT cannot control. All videos used in these experiments are available for download[1]. Figure 6 shows the stages of each experiment and the order of their implementation. Two issues are important: the attention and memory experiments were carried out simultaneously with the same participants, and the formal study on awareness was carried out right after the first condition of the combined attention/memory experiment. Regarding the latter, to generate the stimuli for the formal awareness study, we needed the FRs for each video clip. Conveniently, this is exactly the outcome of the first condition of the attention experiment. On the other hand, the result of the formal awareness study is a single modulation threshold needed to prepare the stimuli for the second condition of the combined attention/memory study. The participants of the awareness studies did not take part in any of the other studies. Thus, there is no risk in interleaving these experiments. Concerning the combination of the attention and memory experiments, all participants received the instructions and performed the memory experiment after completing the attention experiment. The attention experiment evaluates stimulus-based responses and requires that the participant is not given a task. In contrast, the memory experiment was conceived to evaluate a task-based response, requesting participants to remember objects. The duration of the combined experiment was roughly 30 min.

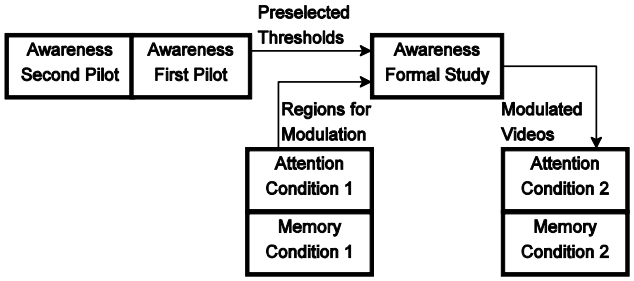

Figure 6. Stages of each experiment in chronological order.

Limitations

The restriction that no faces/hands appear in the clips seems to impact generalizability. However, in our experience, AR applications in non-urban areas easily meet this requirement, and even urban AR applications often involve users looking at equipment without seeing people. We believe the restriction is a reasonable way to control this potential confound now, prior to addressing it in future research.

The results of our experiments do not guarantee that every viewer will attend to and/or remember the selected objects, but that they are more likely to, as compared to the original unmodified condition. However, thresholds can be adjusted interactively by passing a parameter to the SMT implementation (e.g., in response to eye tracker feedback). So, if an application needs to make the effects of the SMT perceptible, it only needs to increase the modulation threshold.

[1] http://hydrosys.icg.tugraz.at/media_files/Saliency

Several factors have been identified that contribute to saliency (e.g., see Wolfe and Horowitz [20]). Of these, the SMT controls contributions in lightness, red–green color opponents, and blue–yellow color opponents, while other factors remain unaffected. In our studies, factors such as motion and size negatively affected results. Future research will need to address how contributors to saliency not controlled by the SMT affect its application. Meanwhile, the effectiveness of the SMT depends on the balance of these factors throughout the input. Avoiding extremes (e.g., small objects) can help in using the SMT successfully.

The main limitation of this approach is registration: how do we decide that a certain portion of the video frame corresponds to a real world object that we want to emphasize. Vision-based object recognition can provide an answer to this question, albeit with limitations of its own.

Outlook

The SMT is implemented in a GPU shader program, and runs at interactive rates on a desktop PC. We are currently experimenting with implementations on an ultramobile PC.

That the modulation is imperceptible is a crucial contribution of our technique. Exit interviews from the attention/memory experiment showed that the mean perceived alteration of videos was equally low for the first (M=2.0, STD=1.214) and second conditions (M=1.7, STD=.801), $t(40)=.922$ $p=.362$ (2-tailed *t*-test). Choosing the right modulation threshold is a key requirement for the SMT to work imperceptibly and effectively. For the experiments presented in this paper, thresholds have only been studied empirically. A thorough study of thresholds for each modulation dimension and their effect on visual attention could greatly improve the selection of modulation threshold.

Our main motivation for developing and experimenting with the SMT is AR, in particular, information-rich visualizations. One idea for mobile devices is to use the SMT as an aid to navigation. We would like to suggest objects related to landmarks and explore whether a navigator would recall having seen them along a path. Our results foster experimentation in this direction. Furthermore, we are aware that the SMT has applications beyond AR; for example, in training, the SMT could be used to suggest that a trainee shift attention towards areas of interest in a scene. A surgeon during training surgery might be reminded of sensitive organs near the work area without visually overlaying any information on the video feed. By varying the modulation thresholds, one could even support using more subtle levels for advanced trainees. Alternatively, physicians following a procedure in real-time could each have the SMT applied to different aspects, depending on a user profile.

CONCLUSIONS

The results of our studies validate the SMT as an alternative means to convey information to the user, *suggesting* attention shifts and *influencing* recall of selected regions without perceptible changes to visual input. These results represent fundamental research and, by no means, cover all the re-

quirements of our motivating scenarios. Still, our experiments address two processes common in HCI: Stimulus-based attention guides the user in the exploration of visual input, playing an important role in tasks such as visual search. Memory is involved in user tasks at several stages (e.g., navigation and visual search). To our knowledge, we are the first to interactively modify videos so that a region we selected contains the highest saliency and experimentally validate its application. We presented three experiments that validate the SMT as an alternative means to convey information to the user. An awareness experiment certifies a modulation threshold that is imperceptible to the user. An attention experiment warrants that regions modulated with the SMT draw a first fixation significantly faster than without modulation. A memory experiment supports that modulation increases recall for selected objects without significant loss in recall for others. In summary, the SMT can significantly shift attention and influence memory to selected areas of a video without the viewer becoming aware of any manipulation. We believe that the results provide sufficient evidence to justify further experimentation in tasks that better match real-world conditions.

The SMT presents an alternative means of attention direction by modifying existing features of the real-world image, instead of adding traditional augmentations (such as pointing arrows or frames). The SMT enables mediated reality, since its premise is modifying the existing video input instead of adding virtual artifacts to it. One advantage inherent to this approach is that it protects context. While the saliency of the context is diminished as that of the focus is increased, the context does not suffer any other degradation. Perceptual issues arising from visual clutter or differences in depth between virtual and real objects are also prevented.

While other approaches to draw attention or influence memory exist, most lack the subtlety of the SMT. The suitability of this technique depends on the application, and also raises ethical issues. In many applications, it will be essential to inform the user that salience modification is being used. Nonetheless, it is our hope that these results provide motivation to contemplate this technique when attempting to design unobtrusive user interfaces.

ACKNOWLEDGEMENTS

This work is partially funded by the EC 7th Framework project HYDROSYS (224416, DG-INFSO). We thank Andreas Duenser for his insight on statistical analysis.

REFERENCES

1. Awh, E., Vogel, E., and Oh, S. Interactions between attention and working memory. *Neuroscience 139*, 1 (2006), 201-208.
2. Bailey, R., McNamara, A., Sudarsanam, N., and Grimm, C. Subtle gaze direction. *ACM TOG 28*, 4 (2009), 1-14.
3. Berg, D. and Itti, L. Memory, eye position and computed saliency. *Journal of Vision 8*, 6 (2008), 1164.
4. Chun, M.M. and Turk-Browne, N.B. Interactions between attention and memory. *Current opinion in neurobiology 17*, 2 (2007), 177-184.
5. Fine, M.S. and Minnery, B.S. Visual salience affects performance in a working memory task. *Journal of Neuroscience 29*, 25 (2009), 8016-21.
6. Itti, L., Koch, C., and Niebur, E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*, 11 (1998), 1254-1259.
7. Itti, L. Models of bottom-up attention and saliency. *Neurobiology of attention 582*, 1980 (2005), 576-582.
8. Itti, L. Visual salience. *Scholarpedia 2*, 9 (2007), 3327.
9. Kim, Y. and Varshney, A. Saliency-guided Enhancement for Volume Visualization. *IEEE TVCG 12*, 5 (2006), 925-932.
10. Kim, Y. and Varshney, A. Persuading Visual Attention through Geometry. *IEEE TVCG 14*, 4 (2008), 772-782.
11. Koch, C. and Ullman, S. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology 4*, 4 (1985), 219-227.
12. Lee, S., Kim, G.J., and Choi, S. Real-time tracking of visually attended objects in interactive virtual environments. *ACM VRST*, (2007), 15-24.
13. Mendez, E., Feiner, S.K., and Schmalstieg, D. Focus and Context in Mixed Reality by Modulating First Order Salient Features. *Int. Symp. on Smart Graphics*, (2010), 232-243.
14. Ouerhani, N., Wartburg, R.V., Hugli, H., and Muri, R. Empirical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis 3*, 1 (2004), 13–24.
15. Rensink, R.A., O'Regan, J.K., and Clark, J.J. On the Failure to Detect Changes in Scenes Across Brief Interruptions. *Visual Cognition 7*, 1-3 (2000), 127-145.
16. Santella, A. and DeCarlo, D. Visual interest and NPR: an evaluation and manifesto. *NPAR*, (2004), 150.
17. Su, S.L., Durand, F., and Agrawala, M. De-emphasis of distracting image regions using texture power maps. *Texture 2005: Proc. 4th ICCV Workshop on Texture Analysis and Synthesis*, (2005), 119-124.
18. Treisman, A.M. and Gelade, G. A feature-integration theory of attention. *Cognitive psychology 12*, 1 (1980), 97-136.
19. Veas, E., Mulloni, A., Kruijff, E., and Schmalstieg, D. Techniques for View Transition in Multi-Camera Outdoor Environments. *Graphics Interface 2010 (GI2010)*, (2010), 193-200.
20. Wolfe, J.M. and Horowitz, T.S. What attributes guide the deployment of visual attention and how do they do it? *Nature reviews. Neuroscience 5*, 6 (2004), 495-501.