DOI: 10.1111/cgf.12092
Eurographics Conference on Visualization (EuroVis) 2013
B. Preim, P. Rheingans, and H. Theisel
(Guest Editors)

Volume 32 (2013), Number 3

Evaluation of Attention-Guiding Video Visualization

K. Kurzhals, M. Höferlin and D. Weiskopf

Visualization Research Center (VISUS), University of Stuttgart

Abstract

We investigate four different variants of attention-guiding video visualization techniques that aim to help users distribute their attention equally among potential objects of interest: bounding box visualization, force-directed visualization, top-down visualization, grid visualization. Objects of interest are highlighted by rectangular shapes and then we concentrate on the manipulation of color, motion, and size. We conducted a controlled laboratory user study (n=25) to compare the four visualization techniques and the unmodified video material as baseline. We evaluated task performance and distribution of attention in a search task. These two properties become especially important when video material with numerous objects has to be observed. The distribution of attention was measured by eye tracking. Our results show that a more even distribution of attention between the objects can be achieved by attention-guiding visualization, compared to unmodified video. Many participants feel more comfortable when they look at bounding boxes and the grid, but improvements in search task performance could not be confirmed.

Categories and Subject Descriptors (according to ACM CCS): H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Evaluation/methodology; I.3.m [Computer Graphics]: Miscellaneous—Video visualization; I.3.3 [Computer Graphics]: Picture/Image Generation—Display algorithms

1. Introduction

Within the last years, a rapidly growing amount of video data from various sources such as video hosting services, CCTV cameras, or animated output from scientific visualizations created a challenge in data analysis. Video visualizations as essential part of multimedia visualization and multimedia visual analytics [CTW*10] represent techniques that can handle the video material as input and visualize important information for its analysis. We consider attention-guiding video visualization techniques as methods that filter raw video data to show important objects of interest. Therefore, they fit in the visual analytics mantra by Keim et al. [KMS*08].

Although an application of the presented techniques to various video material is generally possible, we concentrated our research on video surveillance footage recorded by static cameras. With estimated 40 million surveillance cameras worldwide [Nil09], an immense amount of potential analyzable video data is provided. Attention-guiding video visualization may help observers identify abnormal activities that could be overlooked otherwise, especially in live observations. Surveillance video material is particularly suited for search tasks and can be used to evaluate the techniques with non-experts. Similar to the definition of Henderson and Hollingworth [HH99], we define a video scene as consisting of moving objects of interest and a static background. When users look at surveillance videos, the distribution of their attention may lead to problems. Although recent studies imply that the human tracking of multiple objects is possible [CA05], the number of objects is very limited and the attentional focus has to be shifted to observe scenes with many objects. Since objects need attentional focus for a detailed examination [WRB06], users can miss important changes on objects or even whole objects while concentrating on another one, according to inattentional blindness and change blindness [Ren00]. A uniform distribution of attention could help to reduce these problems. There are various reasons why some objects receive more attention than others. In general, the understanding of visual attention plays an important role in the design process of the visualizations; see the survey of theories of visual attention and visual memory in the context of visualization and computer graphics by Healey and Enns [HE12]. In search tasks, the guidance of attention is

influenced by top-down information, based on knowledge (e.g., "what do I search for?") and bottom-up information provided by different attributes of a stimulus [WBLH03]. The visualization techniques we examine concentrate on the manipulation of bottom-up information.

Our contribution to the field of video analysis is a comparative study of video visualization techniques that achieve a more even distribution of attention in search tasks. *Bounding boxes* and *top-down visualization* as state-of-the-art techniques are compared with *force-directed* and *grid visualization* as new variants in the context of video visualization. In the user study, we used eye tracking to obtain information about the distribution of attention. Although the measurement of eye fixations does not necessarily completely match with attention [Gri96], it provides us with a good impression of the users' viewing behavior.

2. Related Work

Related work in the field of attention-guiding techniques can be found in different fields of visualization. Kim et al. [KV06] modified regional luminance and chrominance in volume visualizations to emphasize regions of interest. By calculating saliency maps based on bottom-up features such as intensity, color, orientation, or motion, attention can be guided by semantic depth of field [ST10] or subtle image modulation [VMFS11]. Hillaire et al. [HLRC*12] provided an attention model for 3D environments that combines bottom-up and top-down features in order to predict human gaze behavior. Bonanni et al. [BLS05] created user-centered interfaces for an augmented reality kitchen with consideration of the user's attentional focus. Biocca et al. [BOTB07] presented an attention funnel to guide the user's attention to objects of interest. Dierker et al. [DMH*09] manipulated the colors of virtual objects to mediate the attentional focus between two partners in a collaborative task. Especially the use of bounding boxes can be found in various applications (e.g., [KCM03]). Schematic rendering methods, similar to the *top-down visualization* can be found in the paper by Girgensohn et al. [GKV*07]. The *force-directed visualization* is adapted from graph drawing algorithms [FR91] but is new in the context of video visualization. However, none of the previous papers provided or evaluated attention-guiding techniques for video visualization.

Kosara et al. [KHI*03] stated the importance of user studies for the evaluation of visualization techniques. By using eye tracking in a user study, various methods help analyze the recorded eye movements. Andrienko et al. [AABW12] provided guidelines for method selection depending on the analysis task. Most of the methods do not provide a meaningful application to dynamic areas of interest. However, different metrics can be used to obtain information about the users' attention. A summary of common eye-movement metrics can be found in the work by Poole and Ball [PB06] and Jacob and Karn [JK03].

3. Perceptual Background

Attention-guiding video visualization techniques manipulate video material to distribute the user's attention evenly among objects of interest. In the design process of a visualization technique, attention-guiding attributes that can be manipulated have to be identified. Wolfe and Horowitz [WH04] provide a list of attributes that might guide the deployment of attention. They distinguish between undoubted, probable, possible and doubtful attributes, as well as probable non-attributes. The undoubted attributes are color, size, motion, and orientation. Two of these attributes and the probable attribute luminance are of special interest for our visualizations:

- **Luminance:** Regarding luminance polarity, the objects and the background often do not differ significantly in unmodified scenes. For most of the practical search tasks, users should mainly pay attention to the objects. The background provides only contextual information, therefore it should be less salient than the objects to avoid distractions.

- **Size:** In a video scene, taller objects might get more attention than the smaller ones. Equalizing objects by adjusting them to a uniform size can lead to new difficulties. Increasing the size can lead to overlap between objects and visual clutter. Shrinking objects can lead to difficulties with the identification of objects and might have to be considered in the context of linear perspective.

- **Motion:** With many different objects in a scene, their motion may differ in many variations of direction and speed. For example, in the video of a traffic scene, a car that goes in the wrong direction should get much more attention than all the cars moving in the right direction. This particular event is very salient and easy to spot, but imagine something happens with one of the cars that are going in the right direction while you are distracted by the one that receives all the attention. By manipulating the object movement, the original information of position and neighboring objects can get lost.

Most of the attention-guiding visualizations mentioned in Section 2 use these features either to find or create regions of high saliency. Due to their importance , we focus our image manipulation on these properties. The following visualization methods were chosen in order to cover different combinations of luminance-, size-, and motion manipulation. The manipulation of color and orientation of objects was considered too distorting. In the visualizations, we prefer the shrinking of objects over increasing their size because the difficulties resulting from overlaps were considered more serious than identification problems of smaller objects. The manipulation of motion is performed by a separation of the scene in two views: a static grid with all objects in equal size and motion; and a scene view to retrieve contextual information. We concentrate on methods that are not only able to guide but also to distribute attention, on condition that important objects are identified and the background is only needed for contextual information.

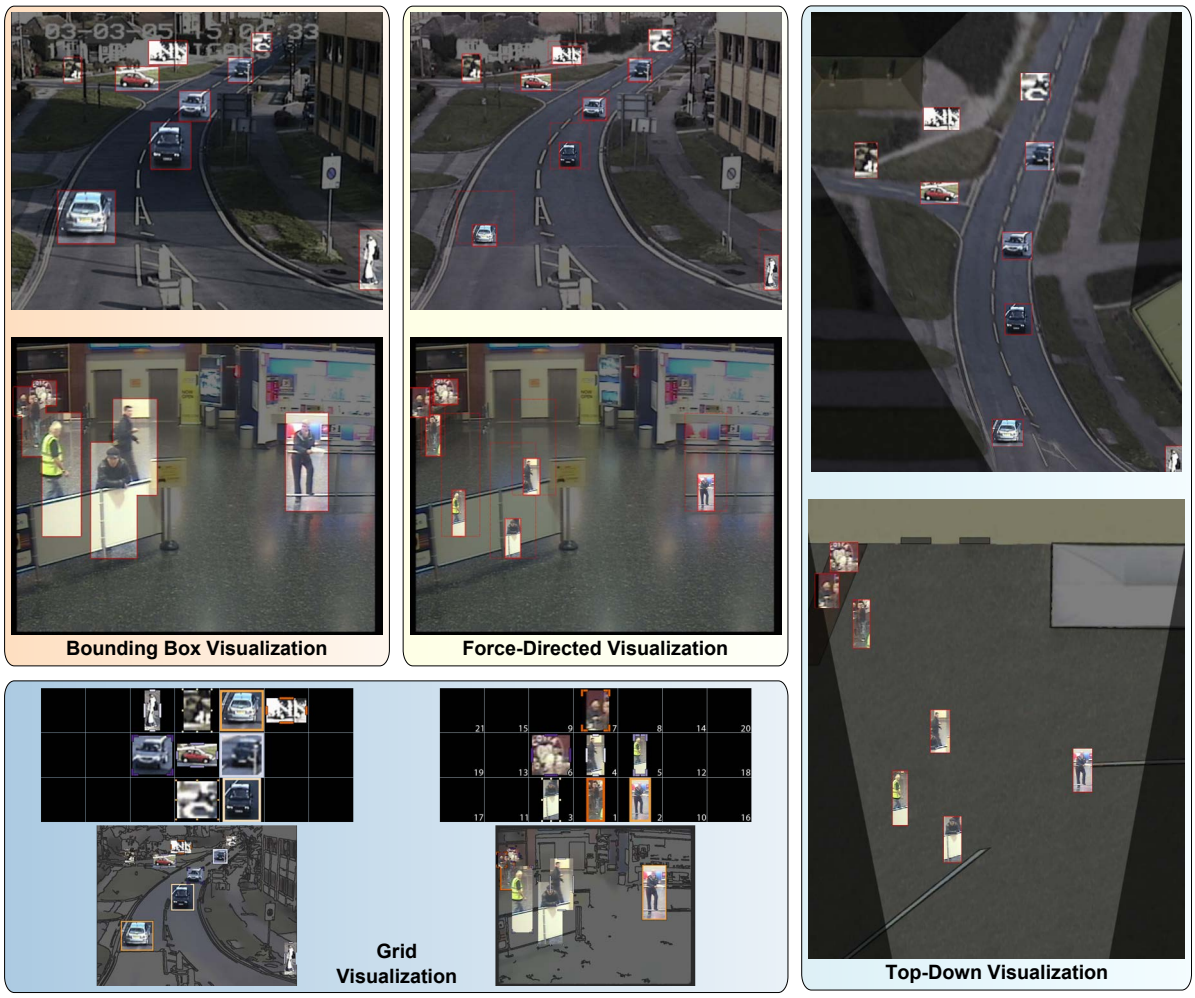

Figure 1: *All attention-guiding video visualizations applied to the i-LIDS dataset for AVSS 2007 [AVS07] (street with cars) and to the i-LIDS multi-camera tracking scenario [i-L10] (area with people).*

4. Attention-Guiding Techniques

We introduce four methods to influence a person's attention. To reduce the saliency of the background in order to draw more attention to objects, the methods reduce either the luminance of the background or exchange them by a static image. To equalize the distribution of attention between objects, different approaches are used: the *bounding box visualization* separates the objects from the background to increase their saliency, the *force-directed visualization* and *top-down visualization* compensate disadvantages of different object sizes by scaling to a uniform size, and the *grid visualization* organizes the also uniformly rescaled objects in a grid to allow the users to shift their focus faster between neighboring objects.

The proposed methods require object recognition and tracking, which could be achieved by automatic computer vision algorithms that are often used in combination with video visualization [HHWH11]. For additional information on state-of-the-art computer vision techniques for video visualization, we refer to Borgo et al. [BCD*12].

4.1. Bounding Box Visualization

A typical approach to highlight objects is the use of bounding boxes. The bounding box visualization surrounds each object in the video by a red rectangle to signalize their importance. Thus, the objects are separated from background and become entities of similar shapes, but different sizes. To prevent visual clutter, the lines of overlapping boxes are removed. To highlight the objects further, the luminance outside the bounding boxes is reduced (see Figure 1).

Since size, movement and appearance of the objects are

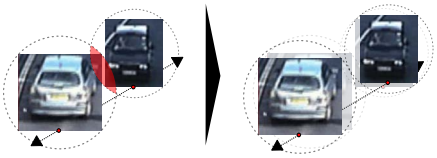

Figure 2: *In the force-directed visualization, overlapping bounding spheres repel each other along the line between their bottom-center points to avoid overlapping.*

Figure 3: *The top-down map is generated by two steps: (i) application of a perspective transformation to the video background and (ii) manual modifications of the transformed image.*

preserved, users do not need additional explanations to understand this visualization. Smaller objects can attract more attention due to object highlighting, but the high level of attraction of large objects is not reduced.

4.2. Force-Directed Visualization

To address the issue that small objects receive less attention than bigger ones, the force-directed visualization is developed. It makes the area size of all objects uniform, where the area size can be adjusted according to the number of currently visible objects in the scene. Although the center-bottom point of the object's bounding box is fixed at its original position, resizing can lead to confusion: objects close to each other in the original scene appear farther away after shrinking. Therefore, a dashed rectangle indicates the original size of the object (see Figure 1).

The different scalings to achieve a uniform size prevent the application of an overlap handling similar to the bounding box approach. As alternative, the method separates the objects by applying a force-directed approach: two overlapping objects repel each other along the line between their bottom-center points (see Figure 2). Additionally, a force is applied that attracts the objects to their original positions. Thus, objects are moving as usual as long as no overlap occurs and begin to dodge each other, otherwise. For smooth dodging, the overlaps of bounding spheres are used instead of bounding boxes. When the distance between overlapping objects is sufficiently large, they move back to their original position.

4.3. Top-Down Visualization

As another approach to reduce overlaps and to provide a good overview in terms of a static background map, the top-down visualization applies a perspective transformation to the scene. Besides of the positional transformation of the objects, they are scaled to a uniform area size chosen appropriately to fit all objects on the map. Objects that occluded each other partially in the original video due to the perspective projection are thus separated in the resulting video and easier to identify as individual objects. Nevertheless, the overlapping bounding boxes show similar images, which can lead to multiple detections of the same event.

For the top-down map representation of the scene, the transformed image can be further edited manually (see Figure 3). Background objects, such as walls and street signs, are removed or replaced by new objects. Generally, the background map may also originate from satellite pictures. To highlight the objects, the background map is depicted with reduced luminance. Additionally, context information for the area outside the field of view of the camera is added with even less luminance. The top-down visualization is also capable of fusing the information from multiple cameras in contrast to the other visualizations.

4.4. Grid Visualization

One drawback of the force-directed and top-down methods is that the uniform size of the objects are smaller in average than in the original video. As mentioned above, this may decrease object identification performance, but increasing their size would lead to numerous overlaps, causing turbulent movements in the force-directed or visual clutter in the top-down method.

For this reason, the grid visualization was developed that trades object size for context information. In this visualization, the scene is split into two components: a grid view and a scene view (see Figure 1). The grid view consists of 21 quadratic cells uniformly scaled for compact object arrangement and is responsible for providing an overview of all objects. By rescaling, the preference of objects depending on their size is equalized. The intention is to place all objects in a compact area to allow users to shift their focus fast between objects and keep all objects near the foveal area. Since context information is not available in the grid view, the scene view provides context by showing the video. The background is simplified by non-photorealistic rendering and reduced in luminance. Since each object in the grid represents an object in the scene view, efficient visual linking of identical objects is crucial. Therefore, the visualization supports an efficient conjunction search for preattentive features [WYS*90] by frames of different colors and shapes. The color scheme consists of 6 diverging colors [HB03] in combination with 4 different shapes, resulting in 24 frames. For scenes with more objects, the method can be extended by defining additional colors or shapes.

5. User Study

We conducted a user study to compare the visualizations in terms of distribution of attention (objective measurement by eye tracking), task performance (objective measurement by task accuracy and reaction time), and a subjective reporting of the most relevant impressions by a questionnaire. In a within-subjects design, we used five video clips from two different scenes in which we added artificial search targets for a detection task.

5.1. Hypotheses

We expect that all the visualizations will influence the participants' attention, compared to the normal video view. The visualizations that equalize the size of the objects, should guide the participants to watch the videos with a more equal distribution of attention. Therefore, we test the following hypotheses:

• **Hypothesis 1 – Distribution of attention:** All visualization techniques should equalize the distribution of attention. Especially by adjusting the objects to uniform size, the top-down visualization, force-directed visualization, and the grid visualization should show a more even distribution of attention than the normal video view.

• **Hypothesis 2 – Task performance:** By drawing attention equally to the objects in the scene, all the visualizations should lead to equal or even better performance in identifying changes on the objects, compared to the normal video view. Especially the grid visualization, which places the objects next to each other to reduce the search area, should make it easier for the participants to detect changes.

We tested the hypotheses in one task by recording eye tracking data of the participants and logging their input with a buzzer during a search task. Additionally, a questionnaire was used to obtain insights in subjective impressions of the participants.

5.2. Stimuli and Task

The visualization type represents the independent variable in our experiment. To measure the attention and task performance, we used five video stimuli. Four parts of the video **V1** [AVS07] with 4:00 minutes each and one video **V2** [i-L10] with 5:15 minutes were presented (both videos with resolution: 720×576 px, 25 fps). We chose two different scenes with a variable difficulty level for our task. **V1** shows a varying amount of activity during playback, including cars, people, and bicycles. The difficulty level to find a particular object is higher than in **V2** due to the size of the scene and the higher number of objects. Therefore, the scene was divided in four parts without the repetition of events. **V2** shows a scene that is less crowded, only people appear and the camera has a smaller field of view. For the evaluation of the visualization techniques, tracking information should be as accurate as possible to eliminate errors in measurement due to annotation noise. Therefore we had to manually create additional ground truth data for the video stimuli. While watching the videos, the participants had to perform the following task:

Figure 4: *The cartoon figure appears several times in a video for a few seconds on existing people or cars. The participants have to find it and hit a buzzer to confirm.*

Change detection task: As Yarbus [TWK*10] already showed in the nineteen sixties, human search strategies while watching scenes depend strongly on the given task. To measure the distribution of attention, the task has to require the participants to constantly shift their focus of attention between all the objects. Therefore, the participants had to find an animated cartoon character that appeared for 4 seconds in place of an existing person or car (see Figure 4). The character was adjusted in size, color, and saturation to fit in the video scenes without being salient. If it appeared on a vehicle, the video was manipulated to look like the character was in the car, looking out of the window or standing on the bumper bar of bigger cars. It represents a clearly defined search target that can be placed at arbitrary times and positions in the video stimuli. The use of that cartoon character has been proved useful for a similar search task in a user study performed by Höferlin et al. [HKH*12]. The character appears 8 times in every video. To confirm a detected character, a buzzer had to be pressed. Since the objects on which it could appear differed in each video, a learning effect for detecting the character was assumed to be negligible.

To counter-balance the experiment, we used a 5×5 Graeco-Latin Square design [CR00]. It combines the visualizations and video stimuli so that every visualization is shown with each stimuli once. By cyclically shifting the columns of the Graeco-Latin Square table, we ensured that each combination appeared once in every position of the watching order. This was done to avoid potential side effects from varying difficulties of the videos. While watching the videos, the participants' eye movements were measured and recorded by an eye tracker.

5.3. Pilot Study

To identify possible issues with the study design, we conducted a pilot study with 7 participants. As a result, we adjusted the time and position of the appearing characters in

every video according to Steven's power law $\psi = k\phi^{0.7}$. With an exponent for visual area [Ste75], the ratio between the mean sensation magnitude ψ of the cartoon character size ϕ and the mean sensation magnitude of the object sizes was approximately equal for each visualization (with $k = 1$). This normalization guarantees that no visualization has an advantage because it would show the character taller than other objects.

5.4. Technical Setup

We conducted the user study in a laboratory with artificial illumination and isolated from outside distractions. The participants were instructed to turn off their mobile phones. They were ordered to sit in front of the eye tracker at a distance of 65 cm. The eye movements were recorded by a Tobii T60 XL eye tracker with a sampling rate of 60 Hz. On the 24" screen of the eye tracker (resolution: 1920×1200 px) the stimuli were presented in the center of the screen with a resolution of 1080 px in height. Depending on the visualization type, the width of the videos was adjusted respectively. To log the user input, we used a buzzer from the game BUZZ!^{TM}.

5.5. Participants

The study was desigend for 25 participants, 7 additional participants had to take part due to insufficient eye position recognition of some participants. A participant's eye tracking data was considered insufficient when more than 25% of the sampled data was discarded by the eye tracker. Six of the participants were female, 19 were male. The average age was 25.3 years; the youngest participant was 19 and the oldest 29. The experiment took 50–60 min, depending on the particular speed of each participant. Each participant was compensated with EUR 10.

5.6. Study Procedure

The participants were asked to fill out a questionnaire that included, among other things, information about their gender and age. Then, we performed vision tests for eye-sight and color vision, to ensure that the participants were physically able to accomplish the given task. Afterward, a tutorial (about 5 min) introduced the 4 visualizations by showing a short video sequence (15 sec), taken from an unused part of **V1**, with each visualization. To explain the task, the cartoon character was introduced first, then a short video (8 sec) with an appearing figure was shown without any visualization. Upon completion of the tutorial, the participants were asked to get in position for the eye tracker calibration. The visualization videos were shown in the order defined by the Graeco-Latin Squares. Between the videos, short breaks were enforced to prevent errors due to fatigue. Recalibration of the eye tracker was performed before every new video. After the task, the participants had to answer a questionnaire.

6. Study Results

We included the results of 25 participants for the evaluation. For statistical analysis, we used non-parametric tests since not all results were normally distributed. For statistical computing, we used the software from the R Project [Rip01]. The results are divided in two sections: the objective measurements concerning the eye tracking data and task performance, and subjective measurement by the questionnaires.

6.1. Objective Measurement Results

6.1.1. Distribution of Attention

We evaluated the distribution of attention by looking at the gaze duration (measured in frames). To determine which object was fixated in a particular frame, we calculated intersections between the bounding boxes of objects and a fovea circle with the current fixation data coordinates as the center to compensate possible errors due to inaccuracy of the eye tracker. Frames with no intersections due to saccades, looking at the background, and missing eye tracking data, were discarded for the calculation. Also, frames containing the cartoon character were not taken into account: due to the task, the users' attention is drawn to the cartoon character on purpose, which would distort our measurement.

We calculate an expectation value E (always greater than, or equal to, one) for each visualization that indicates the expected attention that an object receives. The desired expectation value $E = 1.0$ represents the even distribution of attention. A higher E denotes that the visualization distributes the attention worse, e.g., for $E = 2.0$ it is expected that the objects receive either twice as much or half of the desired attention. To calculate E, we define two attention measures $\tilde{p}_i$ (observed) and $\tilde{q}_i$ (desired) for all objects i in the videos of each visualization:

$$\tilde{p}_i = \sum_{t \in T_i} \frac{g_{i,t}}{h_t} \qquad \tilde{q}_i = \sum_{t \in T_i} \frac{1}{n_t} \tag{1}$$

$\tilde{p}_i$ is calculated by considering all frames T_i including object i, the intersections $g_{i,t}$ with the object, and the number of interesected objects h_t (due to ambiguous object-fovea intersections) in a frame. $\tilde{q}_i$ is calculated by considering all frames that include object i and the number of objects n_t in the corresponding frames. By normalizing $\tilde{p}_i$ and $\tilde{q}_i$ we obtain the probabilities p_i and q_i. Then, E is caculated as:

$$E = \sum_i p_i e^{|\ln(p_i/q_i)|} \tag{2}$$

The logarithm with its absolute value has the purpose to symmetrize the measure: receiving more attention is thus penalized equal to receiving less attention. The inverse of the logarithm (e^x) is applied before calculating the expectation value to preserve the linearity of the measure. The non-parametric Kruskal-Wallis test on the summands of E shows that the visualization type has a significant effect on the distribution of attention ($H(4) = 26.12,\ p < 0.01$). Post-hoc

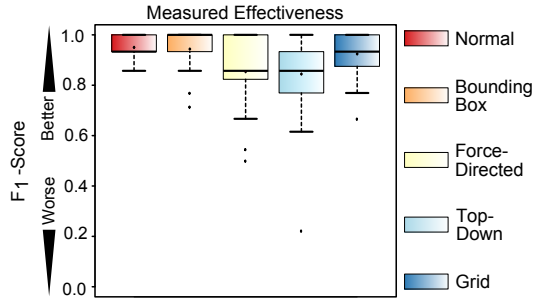

Figure 5: *F_1-score: measured effectiveness for detection of changes (whiskers represent the lowest / highest values within one and a half times interquartile range to the median, the mean is represented by red diamonds).*

pairwise U-tests confirm that, except for the force-directed visualization ($E = 1.551$), the values from the bounding box visualization ($E = 1.522$), the top-down visualization ($E = 1.495$), and the grid visualization ($E = 1.458$) significantly ($p < 0.05$, Bonferroni-corrected) differ from the normal video view ($E = 1.695$). All three visualization techniques show expectation values closer to 1.0. This supports Hypothesis 1, that the participants were able to distribute their attention more equally. Significant differences between the attention-guiding video visualization techniques could not be confirmed.

6.1.2. Task Performance

For an objective measurement of the participants' effectiveness in the search task, we logged their button presses on the buzzer and used this information to calculate scores $F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$ for all stimuli. The results are summarized in Figure 5. A significant effect of the visualization type on the task performance (Kruskal-Wallis: $H(4) = 18.31$, $p < 0.01$) could be confirmed. Post-hoc pairwise U-tests confirm that the top-down method (mean=0.84, sd=0.16) shows significantly worse results ($p < 0.05$, Bonferroni-corrected) than the normal video view (mean=0.95, sd=0.06) and the bounding-boxes (mean=0.94, sd=0.08). Therefore, Hypothesis 2 could not be confirmed for this visualization method. The force-directed visualization (mean=0.85, sd=0.15) shows no significant difference to the normal video view, but the lower mean value could be an indication for identification problems during the task. The grid (mean=0.92, sd=0.08) and the bounding box visualization show no significant differences to the normal video view, indicating that these visualizations had no significant influence on the task performance and did not lead to better results in task performance than the normal video. The resulting F_1-scores were mainly influenced by missed objects and very few false positive reactions.

Additionally, we evaluated the reaction times between the appearance of a character and the corresponding button press by the participants. Although no significant differences could be found between the visualizations, the mean reaction times (in ms) of top-down (mean=1916, sd=911) and force-directed (mean=1904, sd=1021) seem to be slightly longer, compared to the others (e.g. bounding boxes(mean=1574, sd=872)). This could be a result of the smaller object size, which takes the participants longer to identify changes on an object.

6.2. Subjective Measurement Results

To obtain subjective feedback, the participants had to answer a questionnaire after the task. We included six questions that had to be answered for all five visualizations on a 10-point Likert scale. The questions for effort and frustration level were taken from the NASA-TLX Test [Har06], the other four were formulated to evaluate relevant aspects of the visualizations. The supplementary material provides tables containing detailed information about descriptive and inferential statistics of the results. The boxplots in Figure 6 summarize the results.

• **Context: How well could you recognize the spatial context of the objects in the scene?** The visualization type has an significant influence on the rating (Kruskal-Wallis: $H(4) = 49.64$, $p < 0.01$). Post-hoc U-tests reveal significant differences ($p < 0.05$, Bonferroni-corrected) between the normal video and all other visualizations except the bounding box visualization. The normal video was rated best (mean=8.6, sd=1.58) since it represents the scene without any image modification. The bounding boxes (mean=8.2, sd=1.61) seem to have no influence on the subjective impression of spatial context. Force-directed (mean=5.52, sd=2.37) and top-down (mean=6.84, sd=2.34) have problems to preserve the full context, due to resize and displacement. The grid visualization requires to shift attention constantly to the scene view to retrieve context information. The result (mean=3.68, sd=2.66) is significantly worse than the results from the other visualizations, except the force-directed method (mean=5.52, sd=2.37).

• **Relations: How well could you perceive the relations/interactions between objects?** Significant differences (Kruskal-Wallis: $H(4) = 38.75$, $p < 0.01$) were found between the visualizations. All visualizations, except bounding boxes (mean=7.48, sd=2.00), were rated significantly (U-test: $p < 0.05$, Bonferroni-corrected) worse than the normal video (mean=7.96, sd=1.81). Similar to the results of the context question, the grid (mean=3.68, sd=2.44) is significantly worse to perceive relations and interactions than all other visualizations, except the force-directed method (mean=5.24, sd=2.63).

• **Distribution of attention: How equally could you distribute your attention among the objects?** There are significant differences (Kruskal-Wallis: $H(4) = 25.06$, $p <$

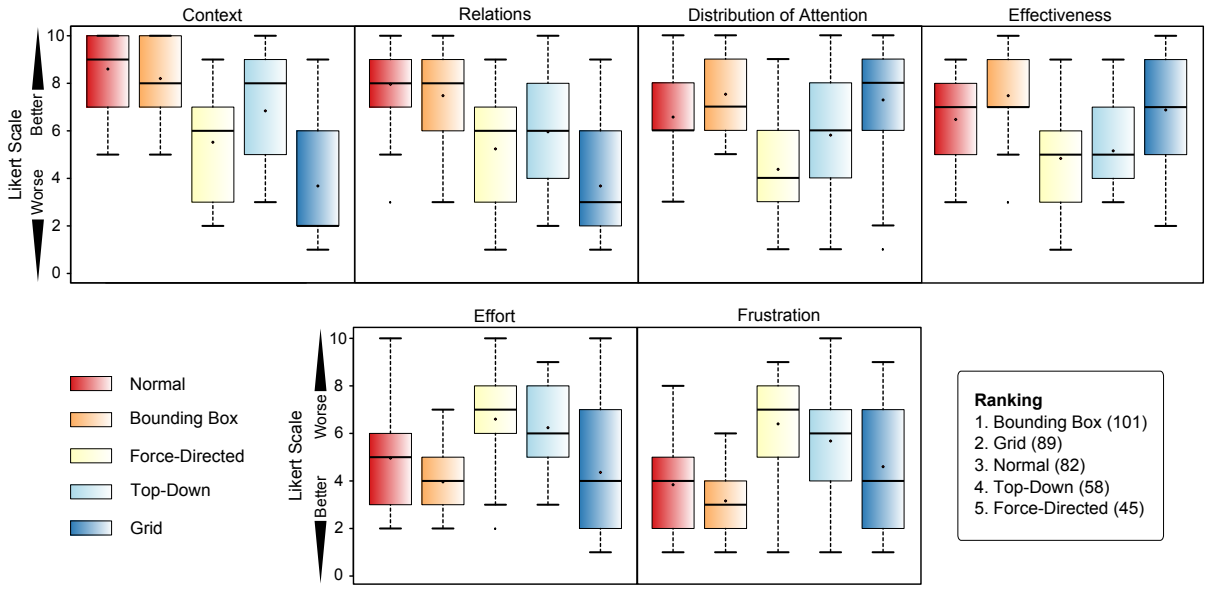

Figure 6: *Boxplots of the results of the questionnaire; bottom right: ranking of the participants for long-time observation (values represent Borda scores).*

0.01) in terms of the distribution of attention. With the exception of the top-down method (mean=5.8, sd=2.55), the force-directed method (mean=4.36, sd=2.2) was voted the worst. The grid (mean=7.28, sd=2.88), the bounding boxes (mean=7.52, sd=1.64) and the normal video view (mean=6.56, sd=2.00) show no significant differences. While watching the video with the grid method, some of the participants preferred to look at the schematic scene below the grid. For them, the grid was more distracting than helpful, which led to lower ratings.

• **Effectiveness: How well did you perform in the given task?** We could find significant differences in the participants' subjective ratings of effectiveness (Kruskal-Wallis: $H(4) = 26.43$, $p < 0.01$). The bounding boxes (mean=7.48, sd=1.76) showed significant (U-test: $p < 0.05$, Bonferroni-corrected) better results than the top-down (mean=5.16, sd=1.77) and the force-directed method (mean=4.84, sd=2.10). The grid (mean=6.88, sd=2.37) was rated significantly better than the force-directed method. Corresponding to the objective measured F_1-score, the subjective measurement confirms that a reduction of the object size in general leads to a decrease of effectiveness during a search task.

• **Effort: How hard did you have to work to accomplish the given task?** The methods show significant differences (Kruskal-Wallis: $H(4) = 27.44$, $p < 0.01$) in terms of effort. The top-down method (mean=6.24, sd=1.81) and the force-directed method (mean=6.6, sd=2.12) were rated significantly worse (U-test: $p < 0.05$, Bonferroni-corrected) than the grid (mean=4.36, sd=2.58) and the bounding boxes (mean=3.96, sd=1.43).

• **Frustration: How insecure, discouraged, irritated, stressed and annoyed were you?** In terms of frustration, significant differences (Kruskal-Wallis: $H(4) = 27.73$, $p < 0.01$) were found between the visualizations. The participants rated the top-down method (mean=5.68, sd=2.23) and the force-directed method (mean=6.40, sd=2.42) significantly (U-test: $p < 0.05$, Bonferroni-corrected) more frustrating than the normal video (mean=3.84, sd=2.12) and the bounding boxes (mean=3.16, sd=1.40). The participants' ratings for the grid (mean=4.6, sd=2.78) diverged in two groups, about 50% rated with a score lower than 4, the others with a score of 4 and higher. According to their comments, some participants were irritated by the object arrangement in the grid.

The participants were asked which method they would prefer if they had to observe a surveillance video over a longer period of time: they had to rank the methods from 1 to 5, beginning with the preferred method. Figure 6 shows the final ranking, based on the Borda count method [LW96]. The bounding box method is the preferred visualization. The grid was ranked second, despite the problems some participants had while watching the visualization.

7. Discussion

The results of the user study revealed some noteworthy facts that we can interpret in the following way:

In terms of distribution of attention, the bounding boxes,

top-down, and grid methods led to significant changes toward a uniform distribution, compared to the unmodified video. This means that these techniques facilitate to distribute the participants' attention more equally. The supplementary material provides heatmaps that support these results. The heatmaps are generated from information of 6000 frames and 5 participants per image. They are normalized by the maximal fixation value of the normal video.

Regarding task performance, we could not find any significant improvements in the participants' ability to detect changes. The top-down and force-directed methods tend to reduce the performance, compared to the normal video. The manipulation of the video material with bounding boxes and the grid did not impair the task. On the one hand, this is a positive result; on the other hand, this means that possible effects of change blindness were not strong enough to decrease the task performance in the normal video. In this case, the techniques with a more uniform distribution of attention could have shown their advantage. One reason for the small differences in task performance could be the task difficulty. Shorter appearances of the cartoon figure could increase the difficulty and help find more significant differences between the visualization techniques. The participants' subjective impressions and comments support the assumption that especially bounding boxes provide the feeling of good effectiveness in the task and the distribution of attention. For a subset of the participants, the grid was also preferred over the normal video, but some participants were irritated by this technique. According to the participants' comments, decreasing the objects' size impaired the possibility to identify changes on objects. Therefore, the top-down and the force-directed method were ranked worst for longtime observations. The object replacement to remove overlaps in the force-directed method was also found irritating by the participants.

We can summarize that bounding boxes and the grid seem to be good methods to draw the users' attention more equally to a larger number of objects of possible interest with a good acceptance by many participants. They feel comfortable to watch these visualizations and would prefer them for longer observation tasks. Nevertheless, object replacement in the grid is still a problem that can lead to irritation. The top-down and force-directed methods seem to be not preferable over the normal video without further improvement. Manipulating the size of objects is an important factor to influence the users' attention. Decreasing the size of objects tends to lead to identification problems in search tasks, while increasing the size of objects leads to visual clutter. A possible trade-off between object size and visual clutter could be to increase only the size of smaller objects until they fit with the larger ones. If objects of interest can be defined and attention should be distributed only among these objects, it should always be the first step to separate them from unimportant information. In cases where this separation is not possible, the use of more subtle methods which perform only slight modifications on the video stimuli could be more appropriate.

8. Conclusion

In this paper, we compared four different visualization techniques for attention-guidance in videos along with the unmodified stimuli. To direct the users' attention to objects of interest, we manipulated the objects in the video stimuli in three aspects: luminance against the background, object area size, and object movement. These manipulations aimed to help users distribute their attention more equally among objects than in the original video. The visualizations were evaluated in terms of task performance and distribution of attention. We combined methods of objective measurement (task performance, eye tracking data) along with a subjective questionnaire. Especially bounding boxes and the grid led to many positive subjective impressions and a more even distribution of attention. Improvements in task performance could not be confirmed.

For future work, we want to refine our task to research if significant differences in task performance can be found at higher difficulty levels. We also want to refine our visualization techniques with the insights we gathered from this study. Since the grid was accepted by many participants, it is interesting to find out what reasons led to refusal by the rest of them and how the use of multiple views can be optimized for the even distribution of attention. As far as simple search tasks are performed, the grid view seems appealing to the participants who used it. Another possibility could be to retain the grid view and replace the scene view by one of the other visualizations. The question how the force-directed visualization can be improved, is of special interest. Further research is planned with different video stimuli and tasks. The application of attention-guiding methods to animated visualizations is an additional research objective.

Acknowledgements

We want to thank Michael Wörner for voice acting. This work was funded by the German Research Foundation (DFG) as part of the SFB 716 / D.5 at University of Stuttgart.

References

[AABW12] ANDRIENKO G., ANDRIENKO N., BURCH M., WEISKOPF D.: Visual analytics methodology for eye movement studies. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2889–2898. 2

[AVS07] i-lids dataset for AVSS, 2007. 3, 5

[BCD*12] BORGO R., CHEN M., DAUBNEY B., GRUNDY E., HEIDEMANN G., HÖFERLIN B., HÖFERLIN M., LEITTE H., WEISKOPF D., XIE X.: State of the art report on video-based graphics and video visualization. *Computer Graphics Forum 31*, 8 (2012), 2450–2477. 3

[BLS05] BONANNI L., LEE C.-H., SELKER T.: Attention-based design of augmented reality interfaces. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (2005), ACM, pp. 1228–1231. 2

[BOTB07] BIOCCA F., OWEN C., TANG A., BOHIL C.: Attention issues in spatial information systems: directing mobile users' visual attention using augmented reality. *Journal of Management Information Systems 23*, 4 (2007), 163–184. 2

[CA05] CAVANAGH P., ALVAREZ G.: Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences 9*, 7 (2005), 349–354. 1

[CR00] COX D., REID N.: *The Theory of the Design of Experiments*. Chapman & Hall/CRC, 2000. 5

[CTW*10] CHINCHOR N., THOMAS J., WONG P., CHRISTEL M., RIBARSKY W.: Multimedia analysis + visual analytics = multimedia analytics. *IEEE Computer Graphics and Applications 30*, 5 (2010), 52 –60. 1

[DMH*09] DIERKER A., MERTES C., HERMANN T., HANHEIDE M., SAGERER G.: Mediated attention with multimodal augmented reality. In *Proceedings of the 2009 International Conference on Multimodal Interfaces* (2009), ACM, pp. 245–252. 2

[FR91] FRUCHTERMAN T., REINGOLD E.: Graph drawing by force-directed placement. *Software: Practice and Experience 21*, 11 (1991), 1129–1164. 2

[GKV*07] GIRGENSOHN A., KIMBER D., VAUGHAN J., YANG T., SHIPMAN F., TURNER T., RIEFFEL E., WILCOX L., CHEN F., DUNNIGAN T.: Dots: support for effective video surveillance. In *Proceedings of the 15th International Conference on Multimedia* (2007), ACM, pp. 423–432. 2

[Gri96] GRIMES J.: On the failure to detect changes in scenes across saccades. *Vancouver Studies in Cognitive Science 5* (1996), 89–110. 2

[Har06] HART S.: NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (2006), SAGE Publications, pp. 904–908. 7

[HB03] HARROWER M., BREWER C.: Colorbrewer. org: An online tool for selecting colour schemes for maps. *The Cartographic Journal 40*, 1 (2003), 27–37. 4

[HE12] HEALEY C., ENNS J.: Attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics 18*, 7 (2012), 1170–1188. 1

[HH99] HENDERSON J., HOLLINGWORTH A.: High-level scene perception. *Annual Review of Psychology 50*, 1 (1999), 243–271. 1

[HHWH11] HÖFERLIN M., HÖFERLIN B., WEISKOPF D., HEIDEMANN G.: Uncertainty-aware video visual analytics of tracked moving objects. *Journal of Spatial Information Science 2* (2011), 87–117. 3

[HKH*12] HÖFERLIN M., KURZHALS K., HÖFERLIN B., HEIDEMANN G., WEISKOPF D.: Evaluation of fast-forward video visualization. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2095–2103. 5

[HLRC*12] HILLAIRE S., LECUYER A., REGIA-CORTE T., COZOT R., ROYAN J., BRETON G.: Design and application of real-time visual attention model for the exploration of 3d virtual environments. *IEEE Transactions on Visualization and Computer Graphics 18*, 3 (2012), 356–368. 2

[i-L10] i-lids multi-camera tracking scenario dataset, 2010. URL: http://www.homeoffice.gov.uk/science-research/hosdb/i-lids/. 3, 5

[JK03] JACOB R. J., KARN K. S.: Eye tracking in human-computer interaction and usability research: Ready to deliver the promises (section commentary). In *The Mind's Eye*, Hyönä J., Radach R., Deubel H., (Eds.). North-Holland, 2003, pp. 573–605. 2

[KCM03] KANG J., COHEN I., MEDIONI G.: Continuous tracking within and across camera streams. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2003), IEEE Computer Society, pp. 267–272. 2

[KHI*03] KOSARA R., HEALEY C., INTERRANTE V., LAIDLAW D., WARE C.: Thoughts on user studies: Why, how, and when. *IEEE Computer Graphics and Applications 23*, 4 (2003), 20–25. 2

[KMS*08] KEIM D., MANSMANN F., SCHNEIDEWIND J., THOMAS J., ZIEGLER H.: Visual analytics: Scope and challenges. *Visual Data Mining 1* (2008), 76–90. 1

[KV06] KIM Y., VARSHNEY A.: Saliency-guided enhancement for volume visualization. *IEEE Transactions on Visualization and Computer Graphics 12*, 5 (2006), 925–932. 2

[LW96] LANSDOWNE Z., WOODWARD B.: Applying the Borda ranking method. *Air Force Journal of Logistics 20*, 2 (1996), 27–29. 8

[Nil09] NILSSON F.: *Intelligent Network Video: Understanding Modern Video Surveillance Systems*. CRC Press, 2009. 1

[PB06] POOLE A., BALL L.: Eye tracking in HCI and usability research. *Encyclopedia of Human Computer Interaction 1* (2006), 211–219. 2

[Ren00] RENSINK R. A.: When good observers go bad: Change blindness, inattentional blindness, and visual experience. *Psyche 6*, 9 (2000). 1

[Rip01] RIPLEY B.: The R project in statistical computing. *MSOR Connections. The Newsletter of the LTSN Maths, Stats & OR Network 1*, 1 (2001), 23–25. 6

[ST10] SU Z., TAKAHASHI S.: Real-time enhancement of image and video saliency using semantic depth of field. In *International Conference on Computer Vision Theory and Applications* (2010), pp. 370–375. 2

[Ste75] STEVENS S.: *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. Transaction Publishers, 1975. 6

[TWK*10] TATLER B., WADE N., KWAN H., FINDLAY J., VELICHKOVSKY B.: Yarbus, eye movements, and vision. *i-Perception 1*, 1 (2010), 7–27. 5

[VMFS11] VEAS E. E., MENDEZ E., FEINER S. K., SCHMALSTIEG D.: Directing attention and influencing memory with visual saliency modulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2011), ACM, pp. 1471–1480. 2

[WBLH03] WOLFE J., BUTCHER S., LEE C., HYLE M.: Changing your mind: on the contributions of top-down and bottom-up guidance in visual search for feature singletons. *Journal of Experimental Psychology: Human Perception and Performance 29*, 2 (2003), 483–502. 2

[WH04] WOLFE J., HOROWITZ T.: What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience 5*, 6 (2004), 495–501. 2

[WRB06] WOLFE J., REINECKE A., BRAWN P.: Why don't we see changes? The role of attentional bottlenecks and limited visual memory. *Visual Cognition 14*, 4-8 (2006), 749–780. 1

[WYS*90] WOLFE J., YU K., STEWART M., SHORTER A., FRIEDMAN-HILL S., CAVE K.: Limitations on the parallel guidance of visual search: Color× color and orientation× orientation conjunctions. *Journal of Experimental Psychology: Human Perception and Performance 16*, 4 (1990), 879–892. 4