Computational and Mathematical Organization Theory manuscript No.
(will be inserted by the editor)

A Computational Model for Human Eye-Movements in Military Simulations

Patrick Jungkunz · Christian J. Darken

Received: date / Accepted: date

Abstract Models of eye movements of an observer searching for human targets are helpful in developing accurate models of target acquisition times and false positive detections. We develop a new model describing the distribution of gaze positions for an observer which includes both bottom-up (salience) and top-down (task dependent) factors. We validate the combined model against a bottom-up model from the literature and against the bottom up and top down parts alone using human performance data on stationary targets. The new model is shown to be significantly better. The new model requires a large amount of data about the terrain and target that is obtained directly from the 3D simulation through an automated process.

Keywords Eye Movements · Human Behavior Modeling · Eye Tracking · Target Detection · Visual Search

1 Introduction

The modeling of target acquisition and detection has always been a major concern for military simulations. In the past, the capabilities of systems were the focus of attention; now the capabilities and the performance of humans need attention. As noted by Evangelista et al (2010), current simulation models

P. Jungkunz
German Naval Office
18147 Rostock, Germany
Tel.: +49-381-8025734
E-mail: patrick.jungkunz(at)gmail.com

C. Darken
Naval Postgraduate School
Monterey, CA 93941
Tel.: +1-831-656-2095
Fax: +1-831-656-7599
E-mail: cjdarken(at)nps.edu

of individual soldiers assume that they search a scene using a fixed pattern, e.g. a sweep from left to right. Anyone who has observed soldiers, especially in an urban environment, surely realizes that this is not an accurate model. Failure to model search accurately results in target acquisition times that are not accurate. Worse, it provides a poor basis for modeling detection phenomena such as false positive detections, i.e. seeing a target where none is present, which can have a significant impact on an operation. Current models of false positive detection can do little better than sprinkle false targets uniformly across the simulated battlefield. If we understood what parts of a scene were challenging for an observer, false targets could be placed in these locations instead.

To improve target detection mechanisms in military simulations, this work proposes to model human eye-movement behavior during target search as a basis for future enhancements in overall models of search and target acquisition. We provide a new model of eye movements and show that it is more accurate than the dominant model in the literature. This model can extract its needed data from a 3D simulation through a process that has been largely automated.

Human visual perception is dependent on the receptive qualities of the retina. The fovea, which is the center of the retina, provides high visual acuity and subtends about 2° of visual angle. This acuity rapidly decreases with higher eccentricity from the center (Rayner and Pollatsek, 1992). The high acuity of the center is necessary for reliable object recognition. It follows that in order for humans to perceive the whole world around them with high acuity they have to perform eye movements. While the gist of a scene can be determined with a single glance, eye-movements allow humans to serially fixate objects in the visual field one after the other to extract high level details from fixated locations Henderson (2003).

This implies that a target can only be detected if the eyes are directed towards that target and attention is deployed to this location. Also, false targets can only be generated at locations fixated with the eyes.

Eye-movements and deployment of visual attention are both necessary to perceive objects (Itti and Koch, 2001a), and they are closely tied to each other (Hoffman and Subramaniam, 1995). According to Itti (2003), there are several factors influencing the deployment of visual attention. These are bottom-up factors, which are visual scene features, for example salient edges or contrasting colors. Visually salient locations in a scene capture attention and the eyes of an observer. In addition to that, there are top-down, task dependent factors driving attention allocation. Humans can voluntarily direct their eyes to locations they want to examine or need to look at based on their current task.

Eye-movement and visual attention modeling is not a new endeavor. One of the best known computational models of visual attention has been described by Itti et al (1998). This model is based on the idea of a saliency map that highlights the locations of a scene that stand out from their background. It has been shown that such salient locations attract the gaze of human observers and that they contribute to the attention allocation of humans (Itti, 2003).

Unfortunately, the model of Itti et al (1998), as well as other state of the art models of visual attention and eye-movements, do not take task dependent information into account. Extensions to this model try to capture some top-down aspects. For example Navalpakkam and Itti (2005) add top-down modulation to the basic model. Top-down modulation refers to the fact that humans are faster to find targets in visual search if they know the target features beforehand. However, this is at best a partial way of capturing task-dependent information.

So far, not a lot of research has been conducted as to how semantically relevant locations influence eye movements. In addition, there is not any visual attention or eye movement model incorporating this type of information. This being said, there is empirical bottom up work on the distribution of fixations of web pages (Buscher et al, 2009) that could potentially be combined with top down task models such as that of Veksler and Gray (2007) to produce a combined model in the spirit of the one presented here. One related model, the contextual guidance model shows that scene features co-occurring with fixated locations can be learned, and these features can be used to modulate a salience map Torralba et al (2006). The major difference between this model and the work proposed here is that the contextual guidance model has to rely on low-level visual features and can only capture task relevant scene locations by learning which low level scene features co-occur with relevant locations. These locations need to be marked manually for the training set and extensive training of the model based on eye-tracking data from human observers is necessary, such that the model can learn the association between target locations and low-level scene features. In contrast, the relevance maps, which will be introduced in this work, capture the meaning of scene locations for the search task directly without having to rely on low-level scene features and without any training.

Previous experiments confirmed that scene elements which one would expect based on first principles to have a meaning for the task are actually examined by viewers. This has been observed on a qualitative basis in the experimental data of Wainwright (2008), and subsequent experiments showed that scene locations with semantic content for the task are prioritized over scene locations which stand out from the background due to their visual features (Jungkunz, 2009).

The model described in the next section describes how semantically relevant scene locations for the task of finding human targets can be captured.

2 Modeling

The eye-movement model described in this work needs a 3-dimensional graphical simulation environment with its underlying geometry as input. This kind of environment is similar to the ones used in first person shooter games, but also in software for military applications which use 3D graphical displays, e.g. the Maneuver Battle Lab (MBL) in Fort Benning, Georgia.

The model that is presented in the following is based on the observation that humans searching for a human enemy target tend to fixate two types of scene locations. First, locations at which a ground soldier could take cover, such as small walls, and vertical edges, such as window or door frames. Second, locations at which a target would blend in well with the environment and would therefore be hard to detect.

The model will capture these two types of locations in a map that highlights the locations with semantic relevance for the search task. Hence, the map is called relevance map.

2.1 Relevance Maps

To capture semantically relevant information from the simulation environment, which is the basis for the relevance maps of the proposed eye movement model, two applications based on the Delta3D game engine are used. These two applications directly operate on a simulation environment which provides the stimuli or scenes for a human observer as well as the input for the eye-movement model. These two applications are the waypoint explorer application and the intervisibility application. The waypoint explorer application (Darken, 2007b) creates a dense hexagonal waypoint mesh which is used in conjunction with the simulation environment by the intervisibility application to create the relevance map.

The Waypoint Explorer Application The waypoint explorer creates the waypoint mesh in the following way. Starting from one or more waypoint seeds, the explorer travels through the simulation environment. It is able to reach every location within the environment which could be reached by a human. Every location the explorer visits is marked with a waypoint. From any location the explorer reaches it tries to step into six different directions by a given step size. The six directions have a regular angular separation of 60 degrees. Thus the resulting waypoint mesh has a hexagonal structure (see Fig.1). The explorer only performs a step if the desired location can be reached by a human. The applications stops when all reachable locations of the simulation environment have been explored. The output of the application is a set of waypoints with its interconnecting links. The model described in this work makes use of the waypoints only.

The Intervisibility Application The set of waypoints and the simulation environment are the input for the second application, the intervisibility application. The output of this program is the so-called pixelbank, which is used to derive the relevance map. For a given observer's viewpoint the application renders a scene, which is an image or a frame of a visual simulation. The image in Fig. 2 shows the simulation environment from the given viewpoint. A scene is rendered once for each waypoint visible from the current viewpoint. Each time, a

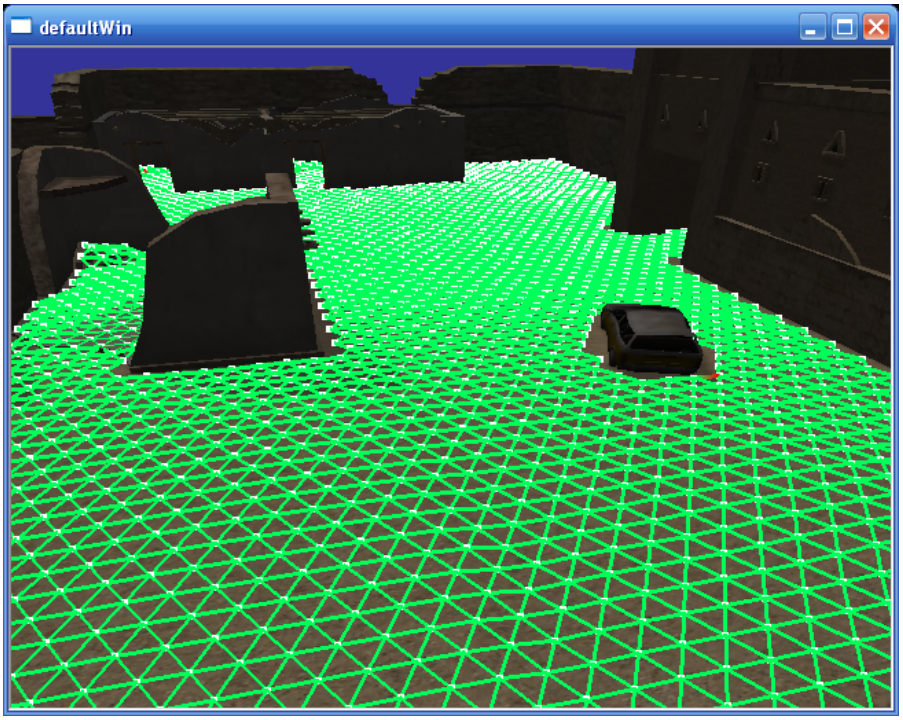


Fig. 1 An example of a waypoint mesh laid out in the environment used in this work. The green lines indicate links between waypoints which can be traversed by a person. The waypoints themselves are located at the intersections of the green lines.

target figure is placed in standing position at a different waypoint before the rendering takes place.

For this target, visibility information is collected, and for every pixel of the target, an entry is made at the respective pixel coordinate in the pixelbank. The pixelbank is a 3-dimensional data structure where the x- and y-coordinates of the pixelbank are image coordinates, i.e., the horizontal and the vertical position in the rendered image or frame of that scene. The z-coordinate of the pixelbank represents the distance of the target from the camera viewpoint in terms of the z-buffer value of that portion of the target. The z-buffer value is a representation of the distance from the viewpoint internal to the computer graphics buffer derived by applying a monotonic function to the Euclidean distance.

The visibility information that is computed for each target pixel and stored in the pixelbank includes the fraction of visible pixels (ratio of pixels visible to an observer to the total number of pixels that would be visible if there were no obstructions) and the contrast of the target to its background. The fraction of visible target pixels can be used to determine locations at which a target can hide behind something. If the fraction of visible pixels is zero, no

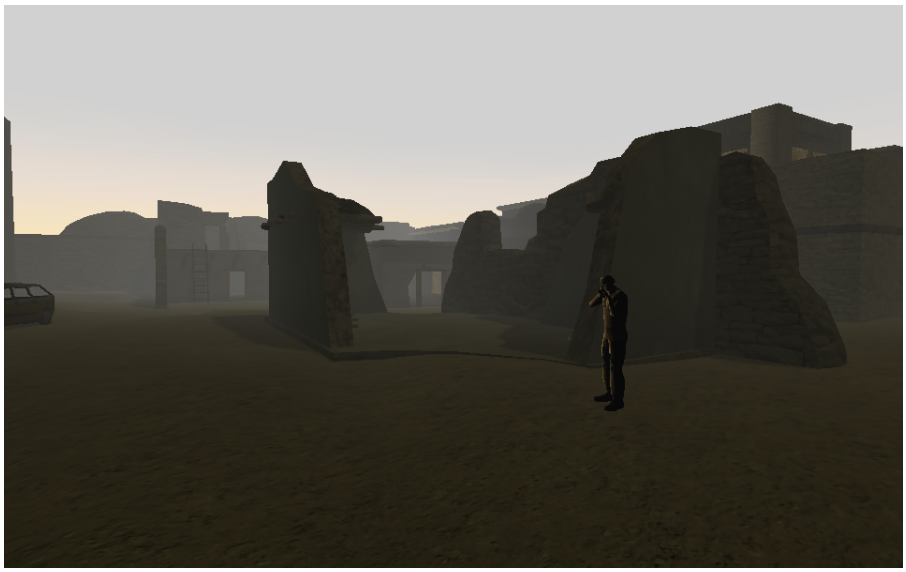


Fig. 2 A scene of the environment used in this work rendered with the target at one of the waypoints. The waypoints are not displayed.

portion of the target is exposed. If it is one, the target is fully exposed. Any number in between indicates that the target is partially covered. The contrast of the target to its background is a measure of the visibility of a target. High contrasts indicate clearly visible targets and low contrasts indicate targets that blend with the background very well. The contrast computation is performed as defined by Darken (2007a). For each color channel, the target 'intensity' for all pixels p of the target is computed using the following formulae:

$$R_T = \frac{1}{n_T} \sum_{p \in T} r^2(p) \tag{1}$$

$$G_T = \frac{1}{n_T} \sum_{p \in T} g^2(p) \tag{2}$$

$$B_T = \frac{1}{n_T} \sum_{p \in T} b^2(p) \tag{3}$$

The background 'intensities' R_B, G_B, and B_B are computed analogously, where the background comprises all pixels within a rectangle around the target, which have a larger scene depth than the target. The rectangle is 5% larger than the smallest rectangle, which would include the target completely (see Fig. 3).

Then, the contrast is computed for each color channel separately:

$$C_R = \frac{|R_T - R_B|}{R_B} \tag{4}$$

$$C_G = \frac{|G_T - G_B|}{G_B} \tag{5}$$

$$C_B = \frac{|B_T - B_B|}{B_B} \tag{6}$$

and the average of the three contrasts is the resulting contrast value.

$$C = \frac{C_R + C_G + C_B}{3} \tag{7}$$

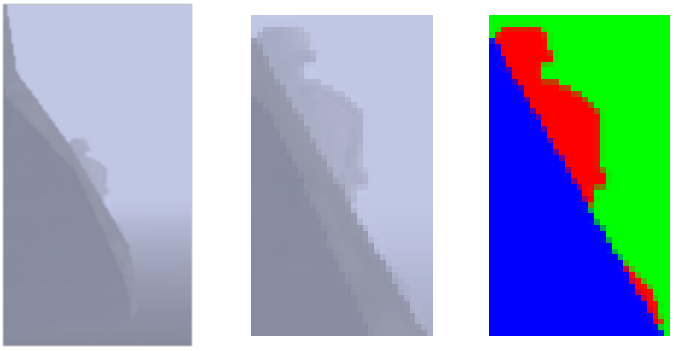


Fig. 3 A target hidden behind a wall (left). A rectangular cutout of the target 5% larger than the smallest rectangle completely including the target (center). The same cutout with the target, background and foreground false colored in red, green, and blue respectively (right).

Two maps are computed from the pixelbank. One map, which is based on the fraction of visible pixels, contains the information about hiding locations. The second map, based on the contrast information, indicates locations at which targets blend in well with the environment.

The hiding location map is derived from the pixelbank by taking the minimum fraction of visible pixels from the list at every pixel. This yields a two-dimensional map ranging from 0 to 1. The width and height of this map are the same as the width and the height of the image rendered from the simulation environment. Pixels with small numbers indicate locations at which at least one target position is occluded and is therefore a likely hiding location. This map is inverted, mapping the range of 0 to 1 to the range of 1 to 0 such that 0 represents a fully exposed target and the numbers close to 1 indicate hiding locations.

Similarly, the contrast map is a two-dimensional map with the same width and height as the hiding location map and the pixelbank. For each x and y image position, the minimum contrast is picked from the pixelbank list at this position. The range of pixel values of this map starts at 0 and can be arbitrarily high. In practice, however, the numbers range from 0 to 1 in most cases. Therefore, all values above 1 are set to one and the result is mapped to the range of 1 to 0. Thus, numbers close to 1 represent locations at which

the target can blend in well with the environment and numbers close to 0 represent locations at which a target stands out well from the background.

The final relevance map is derived by additively combining the hiding location map and the contrast map. Fig. 4 shows an example of a relevance map and Fig. 5 illustrates the derivation of the relevance map from the pixelbank.

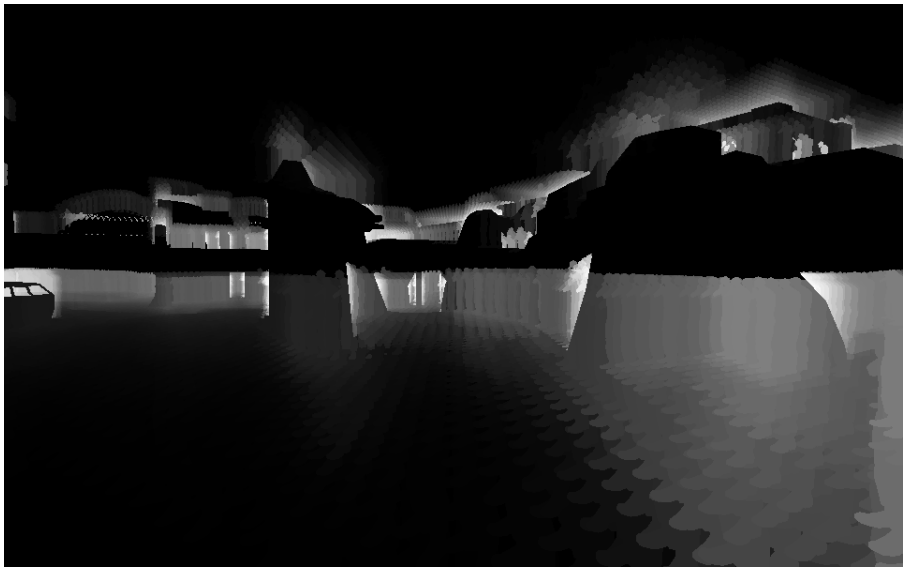


Fig. 4 The relevance map for one scene. White pixels indicate the relevant scene locations.

2.2 Salience Map

Since the control of eye-movements does not only depend on task dependent information, but also on visual scene features, the proposed model includes a salience map in the spirit of Itti et al (1998) as well. The salience map used in this work closely follows the implementation of Itti et al (1998) with a few modifications. Similar to the model of Itti et al (1998) this model considers three basic features: intensity, color and orientation. The details of the salience map computation have been described in Itti et al (1998) and therefore only the changes to the salience map computation will be described here. These changes pertain to the computation of the intensity channel, to the computation of the color center-surround maps and to the normalization scheme used.

The computation of the intensity channel uses the ITU-R 601-2 luma transform to convert the RGB-color values of each pixel into one intensity value.

$$I = 0.299 \cdot r + 0.587 \cdot g + 0.114 \cdot b \tag{8}$$

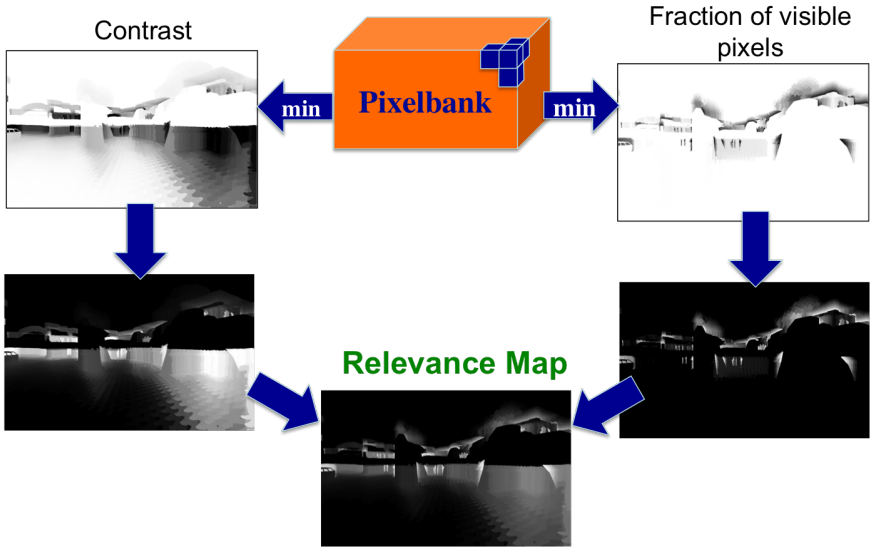


Fig. 5 Derivation of the relevance map from the pixelbank.

This transform takes the different luminance perception of various colors into account.

The implementation of the salience map proposed here follows the suggestion of Frintrop (2006). Instead of using two center-surround channels, four color center-surround maps, one for each color, are used. The computation used to create the basic color feature maps is still as defined by Itti et al (1998).

$$R = r - \frac{g+b}{2} \tag{9}$$

$$G = g - \frac{r+b}{2} \tag{10}$$

$$B = b - \frac{r+g}{2} \tag{11}$$

$$Y = \frac{r+g}{2} - \frac{|r-g|}{2} + b \tag{12}$$

The center surround differences are then computed on six different spatial scales for each color.

$$R(f,c) = |R(f) \ominus R(c)| \tag{13}$$

$$G(f,c) = |G(f) \ominus G(c)| \tag{14}$$

$$B(f,c) = |B(f) \ominus B(c)| \tag{15}$$

$$Y(f,c) = |Y(f) \ominus Y(c)| \tag{16}$$

Where f refers to the fine scale and $c = f + \delta$ to the coarse scale and $f \in \{2, 3, 4\}, \delta \in \{3, 4\}$. The operator $\ominus$ denotes the across scale difference as defined by Itti et al (1998). This means that two maps of a Gaussian pyramid are subtracted from each other. Layer 0 of the pyramid is the original image and the subsequent layers are numbered in ascending order. Before subtraction the coarser map is interpolated to the scale of the finer map.

For every spatial scale, the center surround maps are added up across colors yielding one center surround color map for each spatial scale. These maps are downsampled to scale 4 and added up resulting in the final color conspicuity map. This map is subsequently fused with the intensity and orientation conspicuity maps as defined in Itti et al (1998).

The original bottom-up salience model uses a normalization scheme which is applied to all center-surround maps before being fused into the conspicuity maps of their respective channel. The same normalization is applied to all conspicuity maps before they are combined into the final salience map (Itti et al, 1998). The motivation for normalization is to account for the different dynamic ranges of different modalities and to avoid having locations which are salient in several maps but nonetheless suppressed due to noise in other maps. Different normalization methods have been proposed, but none of them are very convincing (Frintrop, 2006; Itti and Koch, 2001b; Itti et al, 1998). Therefore, an alternate approach is used to compensate for the different dynamic ranges. At first, after basic feature extraction, i.e. after creating the intensity map and the four initial color maps, the maps are scaled from 0 to 1 based on the knowledge that the raw color values range from 0 to 255. Then, each time an operation is applied to a map or several maps are fused, the range of the output is determined by considering the possible range of the input maps and the range the resulting maps could have, based on the applied operator. Next, based on this information the intermediate map is scaled to the range of 0 to 1. If, for example, two maps with minimum values of 0 and maximum values of 1 are added to each other, then the values in the resulting map can range from 0 to 2. This resulting map is then scaled to the range of 0 to 1 again by dividing by 2. The scaling does not depend on the actual values in the map, but on the possible minimum and maximum values a map could have based on the operations performed on the input map up to this point. This ensures, that the ranges of all intermediate maps are confined to the range of 0 to 1, and the final salience map will be in the range of 0 to 1 as well. This mechanism not only ensures that all input maps contribute with equal strength, but also that final salience maps can be compared between images. A map with a green dot on a red background, for example, should have a different salience value at the location of the green dot than a red dot on a background with a slightly different shade of red.

3 Assessing the Model

To assess the quality of the relevance and salience map they will now be compared to eye-tracking data captured from human observers looking for human enemy targets. The data was collected from participants viewing realistic scenes containing one to four targets. These scenes were used to derive the relevance maps as well.

The baseline for assessing the quality of the models are the saliency maps of the Visual Attention model of Itti et al (1998).

3.1 Eye Movement Experiment

An eye-tracking experiment was conducted to record the fixations of human observers looking for a stationary human target.

Participants Nineteen students and faculty of the Naval Postgraduate School in Monterey participated in the experiment after providing informed consent. All participants were members of the U.S. Armed Forces across the four services Army, Marine Corps, Air Force and Navy. The participants were volunteers and did not receive any compensation. All participants were naïve with respect to the hypotheses of the experiment.

Apparatus The stimuli were presented on a 24 inch TFT monitor set to 60 Hz at a resolution of 1920x1200 pixels measuring 52cm x 32.5cm. The stimulus display software was running on a Dell XPS 720 floorstand PC with a Intel Core 2 Quad processor at 2.4 GHz.

Eye tracking was performed with the Seeing Machines FaceLab4 eye tracker. Eye tracking sampling occurred at 60 Hz and the experiment was only conducted for participants for which the screen calibration resulted in a mean error of 1.0° of visual angle or better.

Participants were placed at a viewing distance of 71 cm resulting in the screen covering a visual angle of 40°. The viewing distance was maintained with a modified chinrest used as a chestrest against which participants leaned during the experiment. The head movements were unrestricted.

Stimuli The stimuli presented in this experiment were designed as scenes a ground soldier could possibly encounter in an urban environment. However, all scenes were static, i.e., no movement occurred and all targets were stationary. The targets in the scenes were enemy soldiers in camouflage uniform hiding in structures, behind walls, or other objects in the scene. Enemy soldiers could also be present in open areas. Each scene contained one to four targets. The targets used could appear in four different postures: standing, kneeling, crouching or prone. Sixteen scenes were presented for a maximum of fifteen seconds each. The stimuli can be found in the Appendix[1].

[1] The stimuli have been designed to be displayed on a computer monitor and such that some of the targets are hard to spot. Therefore, the targets in the stimuli are hard to discern

Design and Procedure This experiment was the second part of a series of experiments conducted in one session. After the completion of the first part, the participants continued with this experiment. They were briefed that the scenes of this part of the experiment would be realistic scenes containing one to six instances of the target they already knew from the first part of the experiment. The participants were also informed that the targets could appear in the open or that they could be hiding or taking cover behind other objects, and that the targets could assume four different postures. To familiarize the participants with the possible target appearances, examples of the different postures as well as examples of partially occluded targets were presented. Then, the participants viewed one training scene before starting with the experiment. The scenes were displayed until participants indicated that they had found all targets by saying 'next', but not longer than 15 seconds. Before each scene, a fixation cue consisting of black crosshairs in a white circle on black background was presented to participants. They were asked to fixate the crosshairs until the search scene was displayed.

Although a maximum of four targets were present in each scene, participants were told that there could be one to six targets to avoid search termination based on the number of targets found. Also, the instructions stressed that it was important to find all targets by pointing out that missed targets could be of continuous danger in future.

Before the start of the experiment, the participant's understanding of the task was tested by asking a few questions addressing the key points of the task. After that, the sixteen scenes were presented without any interruption.

Fixation Determination The fixation determination is performed by first finding saccade starting points and end points. Then, all gaze points in between saccades are considered part of one fixation. The fixation location is established by computing the center of gravity of all gaze locations belonging to the fixation. The detection of saccade start and end times is performed using a speed threshold of 8.75° of visual angle per second over two consecutive gaze points. Visual inspection of scene overlays shows that this threshold separates saccades from fixations sufficiently well.

Fixation Maps To compare the participant's fixations with the salience and relevance maps, fixations on one scene over all participants are fused into one fixation map per scene. The fixation maps have the same width and height as the stimuli presented: 1920x1200 pixels. The fixation maps are binary maps containing either values of 0 or 1. Each location of the fixation map for which a fixation was recorded is set to 1. All other pixels of the fixation map are set to 0. This means that a 1 in the fixation map indicates a fixated location and a 0 indicates a location which was never fixated. On average, the percentage of the pixels fixated in a scene is approximately 0.3%.

especially in print. They are best viewed in color. Electronic images of the stimuli can be obtained from the corresponding author.

3.2 Comparison

The fixation maps are compared to the salience and relevance maps using the area under the curve (AUC) of a receiver operating characteristic (ROC) curve following Tatler et al (2005) and Einhäuser et al (2008). An ROC curve plots the false positive rate against the hit rate of a classifier or predictor by scanning through all thresholds applied to that classifier. The hit rate is also referred to as the true positive rate. For the fixation maps, the total number of negative instances for one scene are the number of zeros in the fixation maps, which are all the locations that were not fixated by any participant. Conversely, the total number of positive instances for one scene is the number of ones in the fixation map. These are all the locations that were fixated by at least one participant. The salience maps and the relevance map are treated as predictors of fixations. All values in the map above a certain threshold indicate that this location will be fixated. All values below that threshold indicate that these locations will not be fixated. The locations which are above that threshold and are marked as fixations in the fixation map are hits based on that threshold. All locations which are above the threshold and not marked as fixations in the fixation map are false positives. This assumption is very conservative, since in reality a fixation provides effective viewing for more than just one pixel. Pixels with values above the threshold that are not fixated but lie in the immediate vicinity of the fixation location, will be counted as false positives and not as hits. As a result, the values of the metric used (area under the ROC curve, described in the following paragraphs) will be lower than they should be. However, the proposed comparison metric is still appropriate, since the evaluation of the maps is based on a comparison of the values, not their absolute magnitudes.

Based on the numbers of hits and false positives the false positive and hit rate for a given threshold can be determined and establish one point of the ROC curve. Varying the threshold over the range of the predictor, in this case the salience and the relevance maps (ranging from 0 to 1), yields a set of points forming the ROC curve. A more detailed explanation of the ROC curve creation can be found in Fawcett (2006).

One way of employing the ROC curve to compare classifiers or predictors is to use the area under the curve (AUC). It is easy to determine which of two AUCs is larger. The important thing is, however, that the AUC has a very interesting statistical property. It is equivalent to a Wilcoxcon rank-sum test. This means that the AUC represents the probability with which positive instances can be distinguished from negative instances by using the value thresholded by the classifier in question (Hanley and McNeil, 1982). Applied to the salience and relevance maps, this means that the AUC tells how well these maps correctly distinguish between fixations and non-fixations. Therefore, it is suitable for comparing the predictive power of the bottom-up and top-down maps.

For the comparison of the fixation maps with the predictor maps the eye-tracking error needs to be taken into account. Therefore, the predictor maps are convolved with a Gaussian kernel with the size of 97×97 pixels. This

amounts to 1° of visual angle around every pixel, which is approximately the size of the eye-tracking error. This results in smoothed maps which contain information of the surrounding pixels within 1° of visual angle or 48 pixels at each pixel location.

4 Results

A total of four maps are compared to the fixation maps of each of the 16 scene. This yields one AUC per map and per scene, i.e., 16 AUCs for each map. The ROC curves of all scenes are depicted per map in Fig. 6, and the ROC curves for every scene can be found in the Appendix. The assessed maps are the bottom-up salience map of the original implementation of the model described in Itti et al (1998)[2] (referred to as the Itti map from here on); the re-implemented salience map, which follows the specification of the Itti model with the changes as described in section 2.2, the relevance map and an additive combination of the re-implemented salience map and the relevance map called the combined map. This combined salience/relevance map is computed by adding up the two input maps both weighted with 0.5.

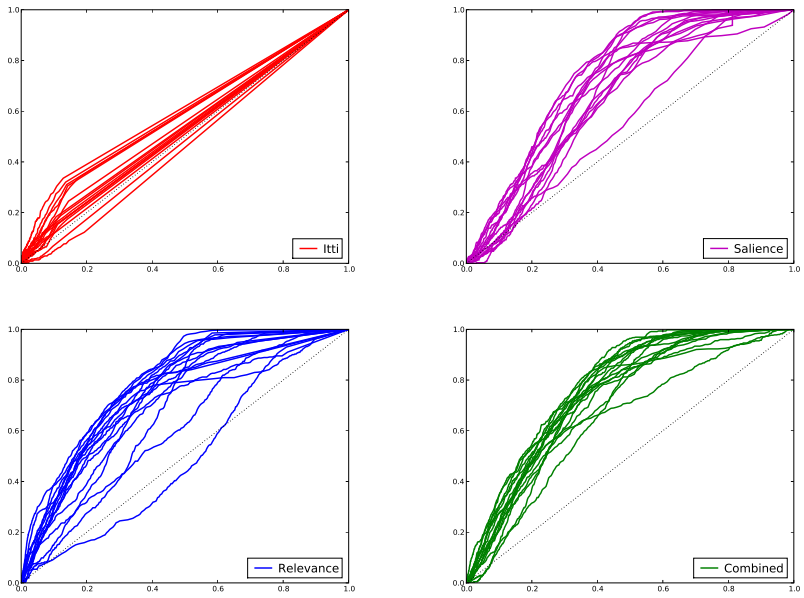


Fig. 6 ROC curves of all sixteen scenes and all four predictor maps. It can be clearly seen how the relevance map and the map combining relevance and salience dominate the pure salience maps.

[2] Implementation derived from `http://ilab.usc.edu/toolkit/downloads.shtml`, last accessed 23JUL2010

The decision to weight the two maps equally is based on a qualitative analysis of the AUCs of all scenes over a range of 9 different weighting factor combinations (0.1/0.9 up to 0.9/0.1). This analysis did not show any indication of improved AUCs when weighting one map stronger than the other. Instead, it could be observed that equal weights for the top-down and bottom-up map resulted in the best performance averaged over all scenes. A more detailed analysis of combination strategies and schemes is left for future work.

To be a useful predictor, the AUC of the maps needs to be larger than 0.5. An area of 0.5 would be achieved by random guessing. The average areas under the curve of the Itti map (μ=0.54, σ=0.04, p=0.0007), the salience map (μ=0.69, σ=0.05, p<0.0001), the relevance map (μ=0.72, σ=0.07, p<0.0001) and the combined map (μ=0.74, σ=0.03, p<0.0001) all statistically significantly exceed 0.5. This means that all of them predict eye fixations better than chance. However, it is apparent that there is a large difference between the average AUCs of the four maps. Therefore, the maps are compared to each other to see if they differ in their predictive power.

The comparison is performed by counting how often each of the maps has a higher AUC, i.e, the number of scenes in which one map outperforms another. The comparisons are based on a sign test using a significance level of 0.05. Comparing the Itti map with the salience map shows that the Itti map is doing better in no scene, and the salience map is doing better in all 16 scenes. The same result is found for the comparison of the Itti map with the combined relevance and salience map. This difference is statistically significant (p<0.0001). As compared to the relevance map, the Itti map is doing better in 1 case and the relevance map in 15 cases. Again, the difference is statistically significant (p=0.0003). Clearly, the Itti map is inferior to all other maps. Looking at the salience map, one can see that it predicts eye fixations better than the relevance map on 4 scenes, whereas the relevance map is a better predictor for 12 of the total 16 scenes. A sign test of this ratio shows statistical significance (p=0.0262). The salience map is also a worse predictor than the combined relevance and salience map. The proportion here is 1:15, which is significant as well (p=0.0003). This means that the salience map performs better than the Itti map only. The other two maps, which both contain information about semantically relevant scene locations, are better predictors of eye fixations than the salience map. Finally, the comparison of the relevance map with the combined map shows that each map is doing better than the other for 8 of the 16 scenes. This proportion is obviously not showing a difference of predictive power (p=0.5). A summary of these results can be found in Table 1.

5 Discussion and Conclusions

The most apparent result of the map comparison is that the Itti map, which is the most well-known model of visual attention allocation and eye movements, is outranked by all other maps. This begs the question of whether the stimuli

Table 1 Comparison of the prediction performance of all maps with all other maps. Each cell shows the fraction of how often the row-map outperforms the column-map over how often the column-map outperforms the row-map. The performance is determined using the area under the ROC curve (AUC). Asterisks indicate fractions which are statistical significant based on a sign test (significance level α=0.05).

	Itti	Salience	Relevance	Salience + Relevance
Itti		0:16*	1:15*	0:15*
Salience	16:0*		4:12*	1:15*
Relevance	15:1*	12:4*		8:8
Salience + Relevance	16:0*	15:1*	8:8	

used for this study are special in some way and not representative of actual environments causing the Itti map to do worse than it would on real world stimuli. Previous research of eye movements on real world photographs using the AUC as a metric as well obtained very similar results (Einhäuser et al, 2008). They report that the Itti map predicts fixations above chance (AUC $>$ 0.5) in 77 out of 93 scenes, which is 82.8% and an average AUC of 57.8% $\pm$ 7.6%. For the scenes in this experiment, the Itti maps predict fixations above chance in 87.5% of all scenes (14 of 16), and the average AUC amounts to 54.0% $\pm$ 4.1%. This means that the performance of the Itti maps in the experiment of Einhäuser et al (2008) is almost exactly the same as the performance observed here.

The most important result of the map comparison is the predictive power the relevance map achieves. The average AUC of the relevance map (71.9% $\pm$ 7.1%) is larger than the average AUC of the salience map (68.9% $\pm$ 4.8%), and the relevance map outranks the salience map on a statistically significant number of scenes. This shows very clearly that semantically relevant scene locations are better predictors of eye fixations than visual salience alone. In addition, the result shows that the novel approach of using information from the simulation environment to determine the semantically relevant locations is highly effective. An even better predictor than the relevance map alone is the combined salience and relevance map. This map outperforms the salience map on 15 scenes and reaches an average AUC of 74.1% $\pm$ 3.0%. This is the expected result based on the assessment of bottom-up and top-down factors on the eye movements in visual search described by Jungkunz (2009) which showed that both visually salient distractors as well as task-dependent influences affect gaze allocation. It is interesting that the combined map does not perform statistically significantly better than the relevance map alone although the average AUC of the combined map is higher than the average AUC of the relevance map.

Looking at the individual scenes more closely reveals that for scenes in which one of the constituent maps has poor performance, the combined map will perform worse than the best constituent map. In cases in which the performance of both maps is rather good, the combined performance increases.

Since the salience map is doing worse than the relevance map for most of the scenes, the salience map can reduce the performance of the combined map as compared to the relevance map alone. In contrast, the contribution of the relevance map to the salience map in the combined map improves performance as compared to the salience map alone.

In other words, there are scenes for which the visual scene features are the governing factor. In this case the salience map predicts fixations better than any of the other two maps.. Then, there are scenes for which the task influence is the governing factor and the relevance map is the best predictor. Lastly, there are scenes, where both visual features and relevant scene information play a significant role, which yields better performance of the combined map than any of the individual maps. The results indicate that in the minority of the scenes, the bottom-up information is the governing factor. In this experiment, there is only 1 of 16 scenes for which the visual information governs the eye movement. This highlights the importance of the semantically relevant scene location over visually salient locations.

In summary, it becomes evident from this research effort that the most influential factor for the prediction of eye fixations is the set of semantically relevant scene locations. In addition, the model presented in this work employs a novel method which allows the direct extraction of semantically relevant information from a simulation environment. This information is fused into the relevance map, which has very good prediction performance.

6 Future Work

Although it is very well known that any kind of movement easily captures visual attention, the effect of moving targets has not yet been included into the described model. Since completely static scenes are rarely encountered in real life, the effect of movement with respect to semantic induced gaze allocation has to be explored.

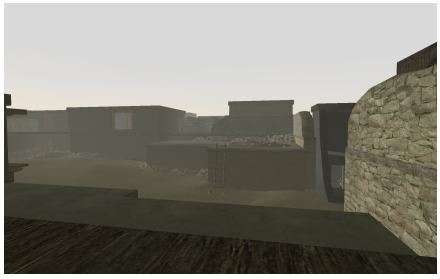
The model described here does not include any knowledge about target features. Previously, Pomplun (2006) has shown that image locations that contain target features receive a higher proportion of eye-fixations than locations which do not. Therefore, it would be interesting to include such a mechanism to see how this changes the prediction performance of the model.

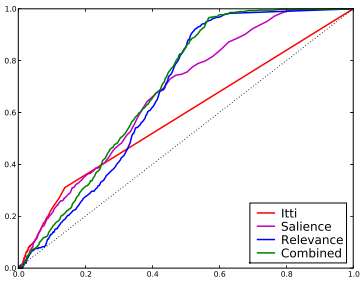
Furthermore, it would be very interesting to explore additional inputs for the creation of the relevance map. At the moment, the relevance map is based on the fraction of visible target pixels and on the contrast of the target to the background. For the contrast input, the size of the target is currently neglected. However, it is not hard to conceive that blending in with the environment is not just a function of contrast, but is also modulated by target size. For example, it would be interesting to explore how a relevance map including the influence 'contrast $\times$ target size' might be constructed, and how the prediction performance of such a map would compare to the currently used maps.

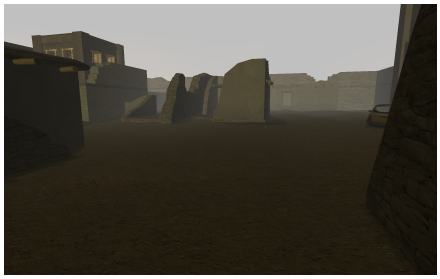
So far, the model has only been assessed with respect to fixation densities. The next step would be to examine fixation order and its relationship to salience and relevance maps. A model for generating fixations including a sequence of fixations could be created based on the following ideas. Currently, the model assigns values to scene locations indicating how strongly a location attracts the gaze in the predictor maps. These maps can be used to generate the scan paths, that is the order of fixations on a scene. Based on earlier findings that one stimulus image elicits very different scan paths for different observers and even for the same observer over different sessions (Mannan et al, 1997), the generation of fixation sequences must not be deterministic. To achieve this, the predictor map is interpreted as a likelihood map. This means each value of the predictor map assigned to a certain pixel is considered to reflect the probability that this particular location is fixated. The initial fixation is determined based on the probabilities assigned to each location. If a location is fixated the likelihood of that pixel and its surrounding area is reduced during the fixation, since a location already fixated is less informative than a location not yet examined. The likelihood of the fixated location decreases over time during that fixation. Thus, the original prediction map is modified over the course of simulated scene viewing. A saccade will occur once the likelihood of the fixated location is less than the likelihood of another location. In addition, based on the observation that human fixations are of limited lengths (Henderson et al, 1999), a cost is associated with the length of the next saccade. To determine the time of the next saccade, this cost is subtracted from each scene pixel based on its distance from the current fixation location. The time of the saccade and the saccade endpoint will be determined based on the probabilities of all pixels of the modified predictor map. The saccade will take place when the currently fixated location does not have the highest likelihood in the map any more. The saccade endpoint will be at the location which does have the highest likelihood at that point in time. The exact details of this model are left for future work.

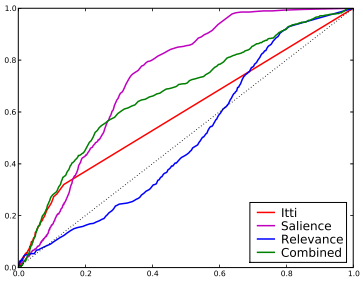
Finally, the model could be extended to not only predict fixations but also to predict target detection probabilities and generate false positives. First of all, it is apparent, that targets which never receive a single fixation will have a detection probability of zero. Furthermore, false positive detections should occur only where a fixation occurred. In addition, the results of the eye-tracking experiment contain false positive predictions. This information can be further analyzed to learn which factors influence false positive generations and detection probabilities.

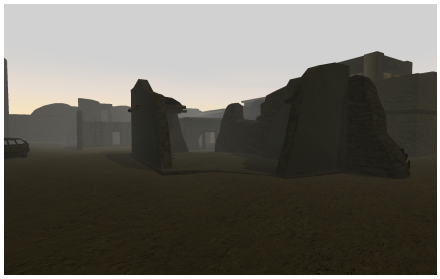
Appendix: Stimuli and Corresponding ROC curves

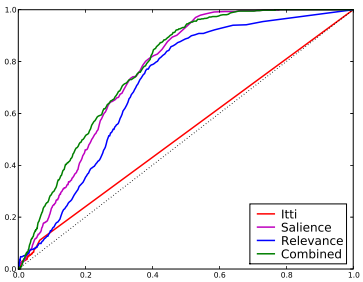


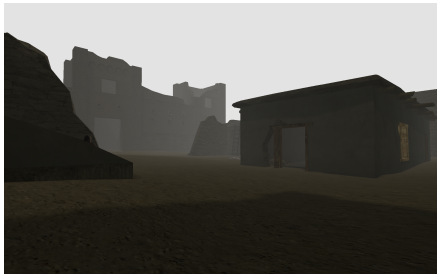


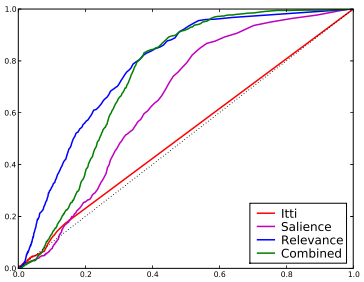


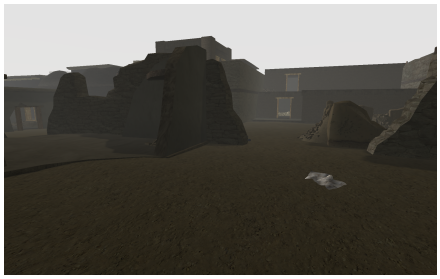


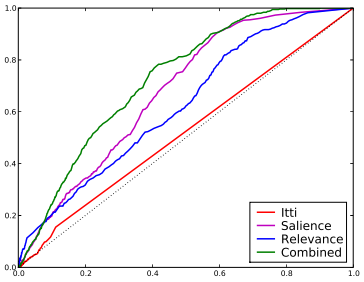


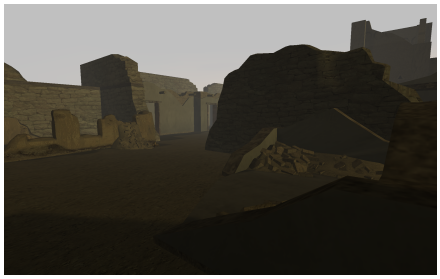


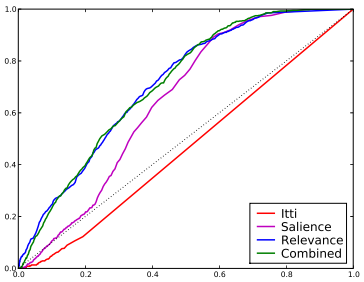


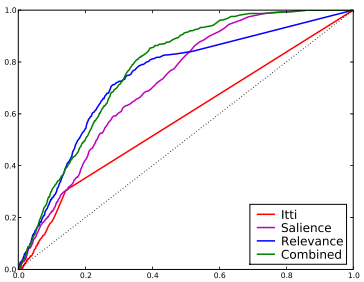

Itti
Salience
Relevance
Combined

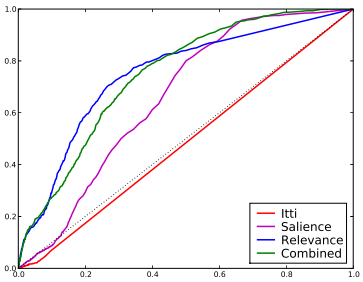


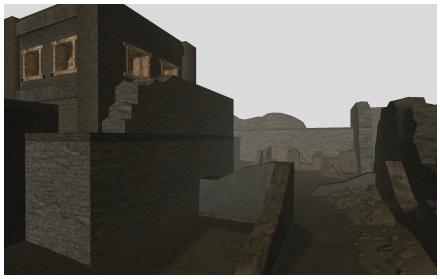

Itti
Salience
Relevance
Combined

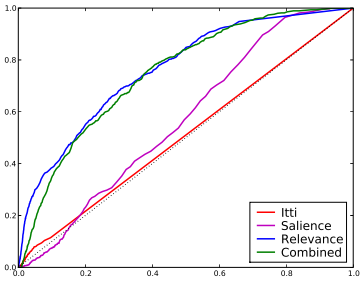


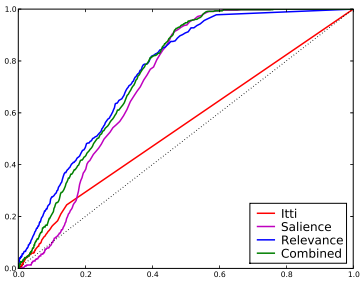

Itti
Salience
Relevance
Combined

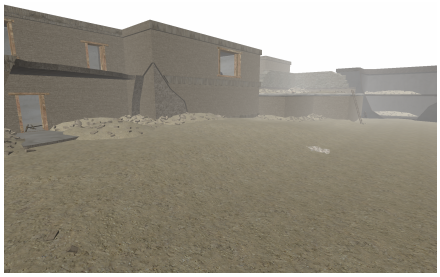

Itti
Salience
Relevance
Combined

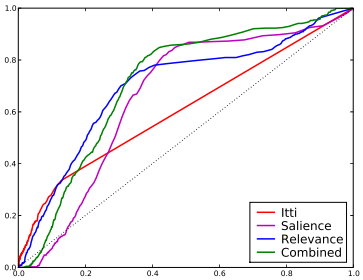

Itti
Salience
Relevance
Combined

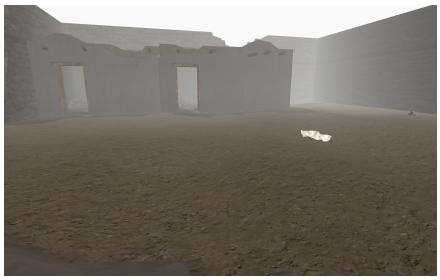


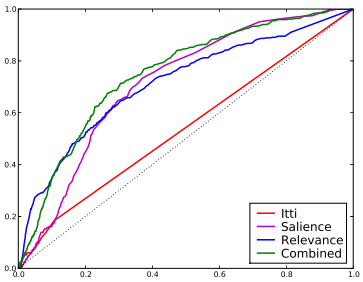

Itti
Salience
Relevance
Combined

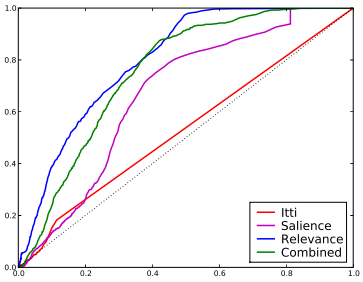

Itti
Salience
Relevance
Combined

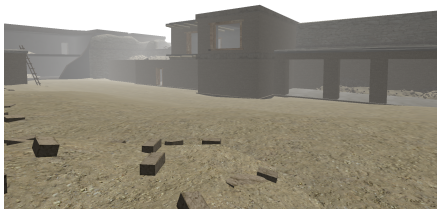


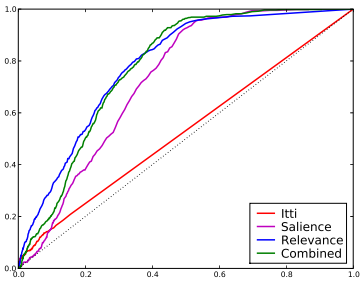

Itti
Salience
Relevance
Combined

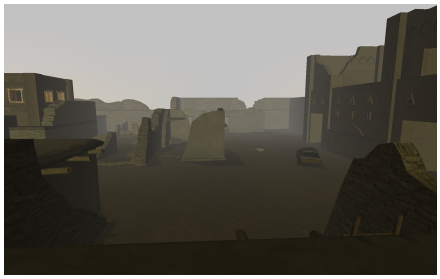


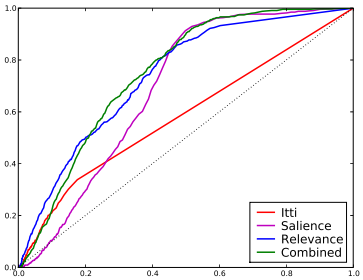

Itti
Salience
Relevance
Combined

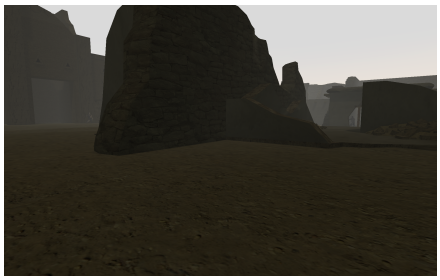

Itti
Salience
Relevance
Combined

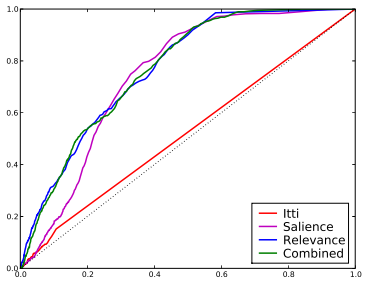



Itti
Salience
Relevance
Combined




Itti
Salience
Relevance
Combined




1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Itti
Salience
Relevance
Combined

References

Buscher G, Cutrell E, Morris M (2009) What do you see when youŠre surfing? using eye tracking to predict salient regions of web pages. In: Proceedings of CHI 2009, Human Factors in Computing Systems, ACM, pp 21–30

Darken CJ (2007a) Computer graphics-based models of target detection: Algorithms, comparison to human performance, and failure modes. The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology 4

Darken CJ (2007b) Level annotation and test by autonomous exploration. In: Proceedings of Artificial Intelligence and Interactive Digital Entertainment (AIIDE).

Einhäuser W, Spain M, Perona P (2008) Objects predict fixations better than early saliency. Journal of Vision 8:1–26

Evangelista P, Darken CJ, Jungkunz P (2010) Modeling and integration of situational awareness and soldier target search. Invited submission to Barchi Prize competition for the 77th MORS

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognition Letters 27:861–874

Frintrop S (2006) VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search. Springer

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36

Henderson JM (2003) Human gaze control during real-world scene perception. Trends in Cognitive Sciences 7:498–504

Henderson JM, Weeks PA, Hollingworth A (1999) The effects of semantic consistency on eye movements during complex scene viewing. Journal of Experimental Psychology: Human Perception and Performance 25:210–228

Hoffman JE, Subramaniam B (1995) The role of visual attention in saccadic eye movements. Perception & Psychophysiscs 57(6):787–795

Itti (2003) Visual attention. In: Arbib MA (ed) The Handbook of Brain Theory and Neural Networks, MIT Press, pp 1196–1201

Itti L, Koch C (2001a) Computational modeling of visual attention. Nature Reviews Neuroscience 2(3):194–203

Itti L, Koch C (2001b) Feature combination strategies for saliency-based visual attention systems. Journal of Electronic Imaging 10:161–169

Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE 20(11):1254–1259

Jungkunz P (2009) Modeling human visual perception for target detection in military simulations. PhD thesis, Naval Postgraduate School

Mannan SK, Ruddock KH, Wooding DS (1997) Fixation patterns made during brief examinaton of two-dimensional images. Perception 26:1059–1072

Navalpakkam V, Itti L (2005) Modeling the influence of task on attention. Vision Research 45(2):205–231

Pomplun M (2006) Saccadic selectivity in complex visual search displays. Vision Research 46:1886–1900

Rayner K, Pollatsek A (1992) Eye movements and scene perception. Canadian Journal of Psychology 46:342–376

Tatler BW, Baddeley RJ, Glichrist ID (2005) Visual correlates of fixation selection: effects of scale and time. Vision Research 45:643–659

Torralba A, Oliva A, Castelhano MS, Henderson JM (2006) Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. Psychological Review 113:766–786

Veksler V, Gray W (2007) Mapping semantic relevancy of information displays. In: CHI 2007 Extended Abstracts on Human Factors in Computing Systems, ACM, pp 2729–2734

Wainwright RK (2008) Look again: an investigation of false positive detections in combat models. Master's thesis, Naval Postgraduate School