Guiding Visual Attention in Remote Operation: Meaning, Task and Object in Post-Disaster Scenarios

Xiaoxuan Zhou[1] and Xing Su[1]

[1]College of Civil Engineering and Architecture, Zhejiang University, Hangzhou, Zhejiang, 310000
xiaoxuan@zju.edu.cn, xsu@zju.edu.cn

Abstract –

In post-disaster scenarios where on-site operations are unfeasible, remote operation of robots or drones by human operators presents an effective and promising solution for survey and search-and-rescue (SAR) missions. These critical missions require human operators to rapidly process extensive visual data and allocate attention within dynamic, complex, and hazardous environments. While previous research has largely focused on the influence of salience and meaning in routine environments, our study shifts this focus towards the unique challenges faced by human operators in emergency settings. Utilizing eye-tracking technology and constructing feature maps of the environment, this research quantitatively assesses the roles of salience, meaning, task demands, and object relevance from the perspective of human operators in post-disaster environments, exploring their interrelationships. Our findings reveal that task demands and object relevance significantly affect how human operators allocate their visual attention, with this influence being modulated by factors such as salience and meaning, while meaning continues to play a predominant role in guiding attention. This study advances our understanding of the visual attention dynamics of human operators in critical SAR missions, providing essential insights for the design of more effective remote operation systems for emergency response.

Keywords –

Visual Attention; Remote Operation; Disaster Management; Eye Tracking

1 Introduction

The increase in severe natural disasters calls for innovative approaches in emergency management[1,2]. Remote search-and rescue (SAR) has been propelled to the forefront of effective emergency response with the advances in robotics and unmanned aerial vehicles (UAVs), offering the ability to access areas otherwise unreachable or dangerous by human responders[3]. Key players in this process are operators who pilot drones or robots in disaster zones. These operators manage equipment and process extensive visual data from the field to guide SAR efforts. The challenge encompasses more than technology operation; it involves analyzing and prioritizing received continuous stimulus. Their effectiveness depends on efficiently allocating attention and identifying crucial information to the visual feed, requiring complex attention skills.

Research into operators' visual attention can elucidate patterns of attention allocation, enabling predictions about focus distribution and, based on these insights, provide tailored support to enhance their operational performance. The study of attention has acknowledged that attention is a limited resource and is influenced by multiple factors[4,5]. Visual attention, as a specific subset of attention research, has been extensively explored, offering valuable frameworks for understanding how individuals process visual stimuli[6]. The framework is divided into two main directions, i.e., visual attention in a scene is driven by bottom-up, low-level image features, such as color, luminance, object feature, and edge orientation, which are combined to form a saliency map[7,8]; and attention is guided by top-down, high-level cognitions, such as knowledge, semantics, and scene context[9,10]. However, because cognition is difficult to represent and compute directly, many current studies on attention guidance still focus primarily on image saliency[11].

However, while these models and theories provide a comprehensive understanding of visual attention in general, they fall short of addressing the specific challenges faced by remote operators in disaster scenarios. The existing literature, while rich, often focuses on controlled environments or tasks that are less complex than those encountered in real-world disaster response. Moreover, the specific aspect of visual attention allocation in interpreting and analyzing real-time imagery, as required in remote operation, has not been extensively studied. Considering the distinctive challenges inherent in disaster environments—including swiftly changing scenarios, profound emotional impacts, and the necessity for prompt decision-making—the discrepancy between extant research findings and practical needs is starkly highlighted. Research has substantiated that variables such as fear emotions[12], visual stimuli[13], and cognitive burdens critically affect both information reception and the distribution of visual attention.

Therefore, this study focuses on post-disaster high-risk task scenarios, aiming to investigate the attention mechanism in the specific environment of remote operations after natural disasters and explore the influencing factors of attention under dynamic and high-pressure tasks. It is expected to provide more efficient and usable solutions for emergency response and disaster management while enriching the theoretical framework of attention allocation modeling.

2 Related works

In the realm of post-disaster remote rescue operations, understanding the visual attention of the driver is critical for effective management and response. The dynamics of visual attention, especially in high-stress and complex environments such as post-disaster scenarios, are multifaceted and are influenced by a multitude of factors ranging from scenario characteristics to individual cognitive processes.

Visual attention is an important component of human visual perception. When confronted with a complex visual scene, human beings will efficiently localize the parts of interest and analyze the scene by selectively processing some regions of the visual input, a process that is also known as prioritized allocation of attention in the presence of limited attention. In order to understand the mechanism of human visual attention, there have been many scholars who have extensively explored the underlying theories of visual attention. Feature-Integration Theory[4], saliency-based visual attention model[7] and graph-based visual saliency model[8] provide essential insights into how visual features are processed and prioritized. These theories highlight the significance of bottom-up stimuli characteristics in directing attention.

With the development of computational vision, many studies have begun to emphasize the effects of top-down high-level features on visual attention. Researches demonstrate that local scene semantics guide attention during natural visual search in scenes[10,11]. Additionally, the work of Võ [14] delves into the deeper layers of scene structure and meaning, further emphasizing the role of high-level cognitive factors in visual attention. This is especially important in disaster scenarios, where quickly understanding the structure of the scene and recognizing meaningful elements in a scene full of debris can be life-saving.

Nowadays, most of the research on visual attention is based on the living environment and simple visual tasks, but the tasks and environments in remote rescue operations are a new dimension and challenge for visual attention research. Fan, Li, and Su[15] discuss the construction of human visual attention maps in teleoperation, pertinent to remote rescue scenarios where operators navigate through rubble remotely. The augmentation of reality, as explored by Eyraud, Zibetti, and Baccino[16], demonstrates how technological enhancements can alter the allocation of visual attention, a factor critical in designing remote rescue operation interfaces.

Understanding visual attention in the context of post-disaster remote operations has practical implications. Driewer, Schilling, and Baier[17] discuss human-computer interaction in rescue systems, highlighting the need to tailor these systems to accommodate the visual attention patterns of operators. Additionally, Rea et al.[18] emphasize the importance of attention-grabbing techniques in robot teleoperation, a key component in remote rescue scenarios.

In this study, we synthesize multiple factors such as salience, meaning, scene, and object semantics to explain and predict patterns of human attention in non-routine scenarios. The study delves into the factors influencing operators' attention in post-disaster remote search and rescue, with a special focus on the particular mechanisms of attention in high-risk tasks and continuously changing scenarios.

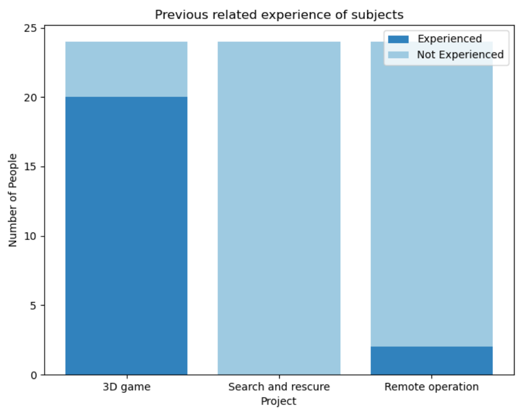

Figure 1. Previous related experience of participants

3 Method

3.1 Eye movements: an experiment in post-disaster remote SAR

We employed a rigorously designed experimental framework utilizing VR technology. A virtual 3D scene of a two-story building after an earthquake was built in

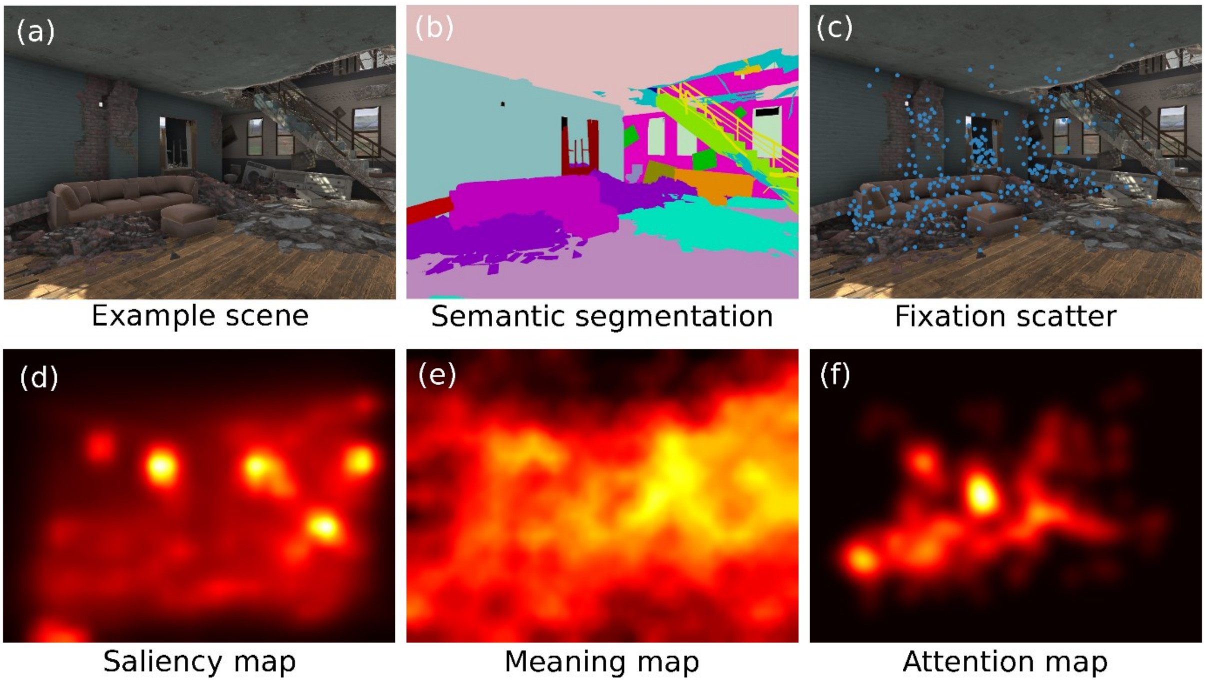

Figure 2. Maps for an example scene. (a) Scene (b)Semantic segmentation mask (c) Fixations scatter plot (d)Saliency map (e)Meaning map (f)Attention map

Unity as a stimulus for the experiment. Users were able to roam around the model following a specified route while completing tasks related to search and rescue. Recognizing the critical role of stress in such high-stakes environments, we integrated time constraints and multitasking requirements to induce realistic stress levels. A pilot test with two participants was conducted before the formal experiment to verify the appropriateness of the experimental design and ensure that the system functions properly.

3.1.1 Participants

A total of 24 (12 males and 12 females) college students volunteered to participate in this experiment. The average age was 23.1 years old. As Figure 1, none of the participants had experience in remote or on-site search and rescue, ensuring a baseline level of expertise consistent across the sample. However, most of the participants had experience with 3D gaming and teleoperation, reflecting a proficiency with virtual environments. This participant profile was intentionally selected to elucidate the intrinsic characteristics of visual attention in SAR scenarios, absent of specialized training biases. Such a selection criteria facilitate the extrapolation of our findings to a broader audience, potentially enhancing the inclusivity and effectiveness of remote SAR training programs.

Table 1 Categories of objects

Object type	Object class
Danger	broken wall, concrete floor, cooking bench, droplight, fallen wall, light, picture, rebar
Information	bed, broken bricks, broken rubble, broken door, closet, door, stairs, toilet, window
Environment	background, bedside table, book, cabinet, scattered ground object, chair, cupboard, kitchen hood, curtain, ventilation, grid, fridge, tableware, shelf, sofa, table, wall, washbasin, washing machine, wooden floor

3.1.2 Scene design

The stimulation scene of the experiment is the indoor environment of a two-story building after an earthquake developed in Unity modeled after a real scene (Figure 2(a)). The model was meticulously designed to closely mimic post-disaster scenarios, facilitating immersive task execution by the participants. Considering the post-earthquake SAR mission, the building's attributes were set as a residence, and the interior design was arranged according to common home life scenes, including living room, kitchen, bathroom, bedroom, and other spaces. In the scene, various kinds of rich objects were arranged to enhance the realism of the scene, and based on this, after removing some object classes that were small in size and could be recognized as having no impact on the study, the points of interest were formed based on the object classes as a unit. As shown in Table 1, the objects were categorized into three categories (danger, information, and environment) according to the task.

3.1.3 Apparatus

Eye movements were recorded using a Tobii Eye Tracker 4C at a sampling frequency of 90Hz. The tracking accuracy of the eye tracker was 38° and 29° horizontally and vertically, respectively. Participants were seated 85 cm from the 21-inch screen, giving the scene a viewing angle of approximately 26.5° × 20° at 1024 × 768 pixels.

3.1.4 Experimental procedure

The experiment was structured into four distinct phases. Initially, participants were provided with a standardized guide to familiarize themselves with the experimental scenario, roles, and tasks. Main tasks entailed identifying lifeforms, ensuring UAV and future personnel safety, and learning about the indoor environment for later tasks. Operators viewed a video of the urban road scenario for task understanding, with 5 minutes allocated for mastery.

Next, participants underwent a 5-point calibration with an eye-tracker for accuracy. In the second stage, they practiced the task for 1 minute. The third stage, the main test, lasted 15 minutes, featuring an automated scene change along a fixed route with static intervals for detailed observation. Throughout the route, there were 30 designated pause points where participants were required to identify and click on objects necessitating attention following task comprehension, all within an 8-second window at each location.

The final stage involved completing a questionnaire about the experiment and gathering essential information.

3.2 Analysis

3.2.1 Data processing

For the eye movement data, gaze and sweep were distinguished by an Initial-Velocity Threshold (I-VT) filter, with the threshold set at 30°/s. The I-VT filter was based on the assumption that if the eye movement data exceeded the threshold, then the eye movement was most likely intentional, otherwise it could be recognized as noise or other non-intentional eye movements. In addition, any gaze shorter than 60 ms and longer than 1500 ms was excluded as an outlier (Figure 2(c)).

For the experimental scenarios, each type of object was replaced in unity with a specific color material without illumination to form a 3D scene with a semantic segmentation mask (Figure 2 (b)).

3.2.2 Attention maps

Attention maps were generated as described in Henderson and Hayes[9]. Briefly, a fixation frequency matrix based on the locations (x, y coordinates) of all fixations was generated across participants for each scene. a Gaussian low-pass filter with a circular boundary and a cutoff frequency of - 6 dB was applied to each matrix, to account for foveal acuity and eye-tracker error. The spatial extent of the low-pass filter was 152 pixels in diameter (Figure 2(f)).

In addition, the significance and saliency maps were normalized to a common scale using image histogram matching using the gaze density maps of each scene as a reference image before statistical analysis.

3.2.3 Visual saliency maps

The saliency maps(Figure 2(d)) for the 30 stops in the scene routes were computed using the Graph-Based Visual Saliency (GBVS) toolbox with default settings[8]. GBVS uses information about the global structure of the image to improve the efficiency of saliency computation. is a well-known saliency model that performs well in complex scenes.

3.2.4 Meaning Maps

Meaning maps of 30 stops in a scene route were created according to the method proposed by Henderson and Hayes.

The stimuli of the session consisted of images of the 30 stopping points. Each image was segmented into circular patches on two scales: fine (300 patches/image) and coarse (108 patches/image.) There were ultimately 9000 unique fine patches and 3240 unique coarse patches.

A total of 132 participants rated the patches on the web application we built. All participants were university students from Zhejiang University, and each of them was allowed to participate in this experiment only once. Each participant would rate 300 randomized patches after reading the experiment content and the two patch examples of low meaning and high meaning. Participants were asked to rate each patch on a 6-point likert scale (very low, low, somewhat low, somewhat high, high, and very high) based on their understanding of the degree of meaning of each patch. Each unique patch was rated by at least 3 independent raters. Finally, the meaning map of each stopping point image was obtained by averaging, smoothing, and combining the ratings of both segmentation scales (Figure 2(e)).

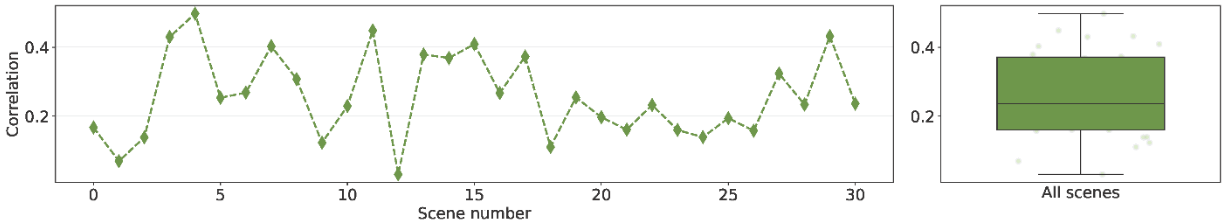

Figure 3. Correlation between meaning and saliency maps

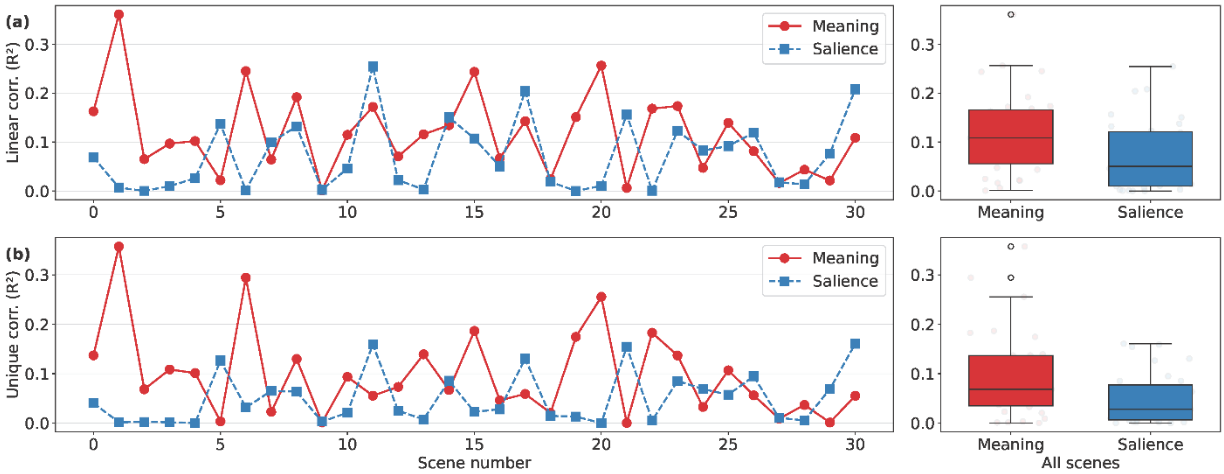

Figure 4. Squared linear correlation and semi-partial correlation by scene

4 Results

4.1 Meaning and visual saliency

Following Henderson and Hayes[9], we used squared linear and semi-partial correlation to analyze meaning maps and saliency maps across the whole scene.

It is suggested that meaning and visual salience are highly correlated across scenes and that the guidance of attention by visual salience may come from meaning. This argument has been validated to some extent in everyday scenarios after the significance map approach was proposed. **Error! Reference source not found.** shows the correlation between meaning and salience for each scene in a complex environment after a disaster. On average, the correlation is 0.26 (s.d. = 0.12) across the 30 scenarios. A one-sample t-test confirmed that the correlation was significantly greater than zero, t(29) = 11.7, $p < 0.0001$, 95% confidence interval (CI) [0.21, 0.30]. This demonstrates that meaning and visual salience are correlated even in non-daily scenarios. It is important to consider the relationship between these two when investigating their role in visual attention modeling.

Figure 4(a) presents the correlation of meaning and salience on attention for all scenes. For the mean squared linear correlation across the 30 scenes, meaning explained 12% of the change in attention (M = 0.12, s.d. = 0.08), while salience explained 7% of the change in attention (M = 0.07, s.d. = 0.07). A two-tailed t-test showed that this difference was statistically significant, t(58)=5.63, $p < 0.0001$, 95% confidence interval [0.004, 0.084].

The ability of meaning and salience to independently explain attention was further explored by computing squared semi-partial correlations, controlling for shared variance in attention. Figure 4(b) illustrates the unique variance of meaning and salience on attention across all scenarios. On average, meaning independently explained 10% of the variance in attention (M = 0.10, s.d. = 0.09), double the ability of salience to explain it (M = 0.05, s.d. = 0.05). This suggests that meaning, controlling for salience, produces 10% additional variance in the attention graph; whereas salience only produces 5%. This difference remained significant by a two-tailed t-test (t(58) = 2.56, $p < 0.01$, 95% CI [0.01, 0.08]). This suggests that meaning relative to visual salience continues to dominate directing attention in cluttered post-disaster scenes.

4.2 Obejct value

Some studies[19,20] have demonstrated that objects predict the allocation of attention points in non-search

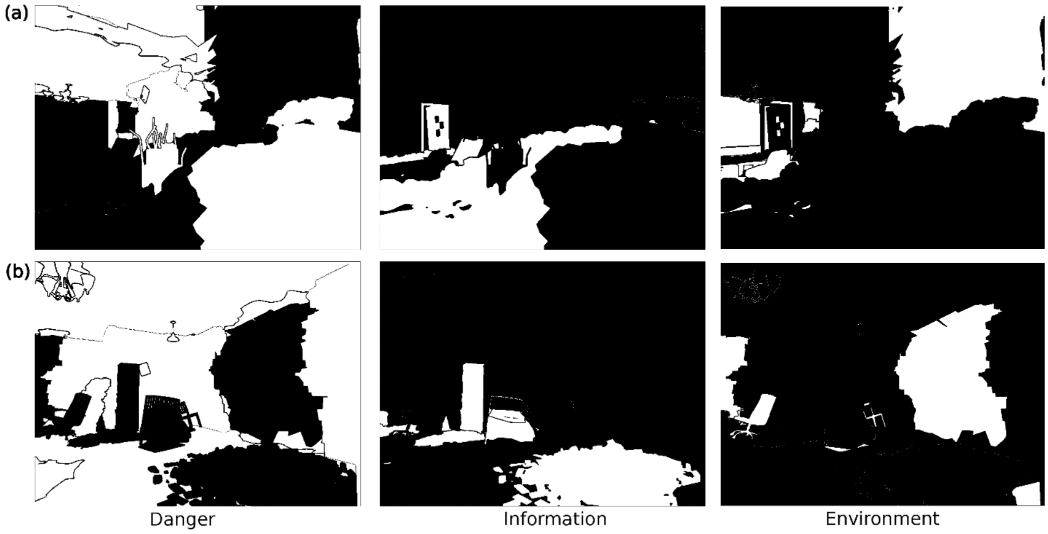

Figure 5. Object semantic category maps of two example scenes

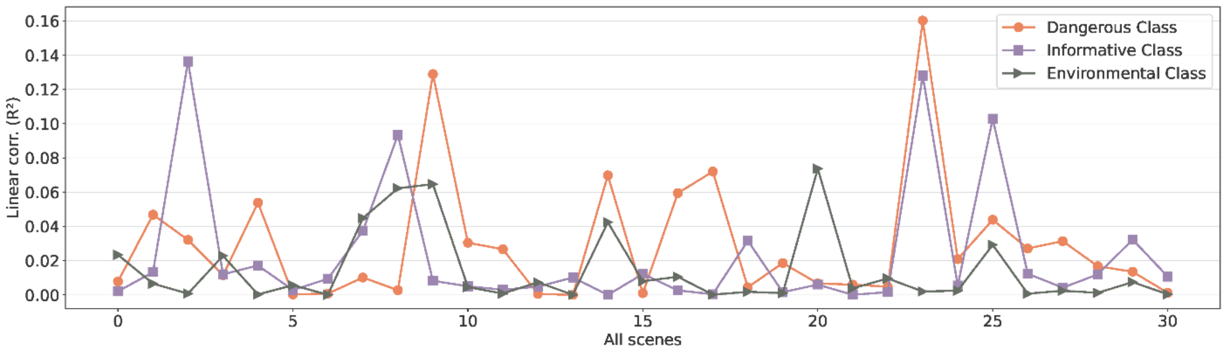

Figure 6. Squared linear correlation between object semantics and attention

tasks. That is, when performing a non-search task, people's attention points tend to be drawn to specific objects in an image.

In the experiments, task and object semantics were highly correlated. As in Figure 5, segmenting the object semantics and distinguishing labels according to different categories (belonging to the category and not belonging to the category) forms an attribute map of the object semantics.

In visual search, attention may be guided by the target or by distractors. Figure 6 shows the squared linear correlations between different semantic categories of objects and attention in all scenes. The danger category explained 2.9% of the variation in attention (M = 0.029, s.d = 0.04), with a one-sample t-test significantly greater than 0 ($t(29) = 4.36$, $p < 0.001$, 95% CI [0.016, 0.043]); the information category explained 2.3% of the variation in attention (M = 0.023, s.d. = 0.04), with a one-sample t-test significantly greater than 0 ($t(29) = 3.43$, $p < 0.01$, 95%CI [0.009, 0.037]); and the environment category explained 1.4% of the attentional change (M = 0.014, s.d. = 0.02), with a one-sample t-test significantly greater than 0 ($t(29) = 3.73$, $p < 0.001$, 95%CI [0.006, 0.022]). This result demonstrates that object semantics guided visual attention in the task and that the target objects (hazard class, information class) were able to play more of a role relative to distractors (environment class), achieving a more task-appropriate guidance.

Figure 7 illustrates the results of the linear correlation analysis of semantic attribute maps with meaning. For all scenarios on average, the correlation coefficient between the danger category and the meaning map was -0.14 (M = -0.14, s.d. = 0.20), with a significant one-sample t-test ($t(29) = -3.80$, $p < 0.001$, 95% CI [-0.217, -0.065]); the correlation coefficient between the information category and the meaning map was 0.076 (M = 0.076, s.d. = 0.15), with a significant one-sample t-test ($t(29) = 2.76$, $p < 0.01$, 95% CI [0.020, 0.132]); the correlation coefficient of the environmental category with the significance map was -0.018 (M = -0.018, s.d. = 0.146) and the one-sample t-test was non-significant ($t = -0.6712$, $p = 0.507 > 0.05$). This suggests that the target objects (danger category,

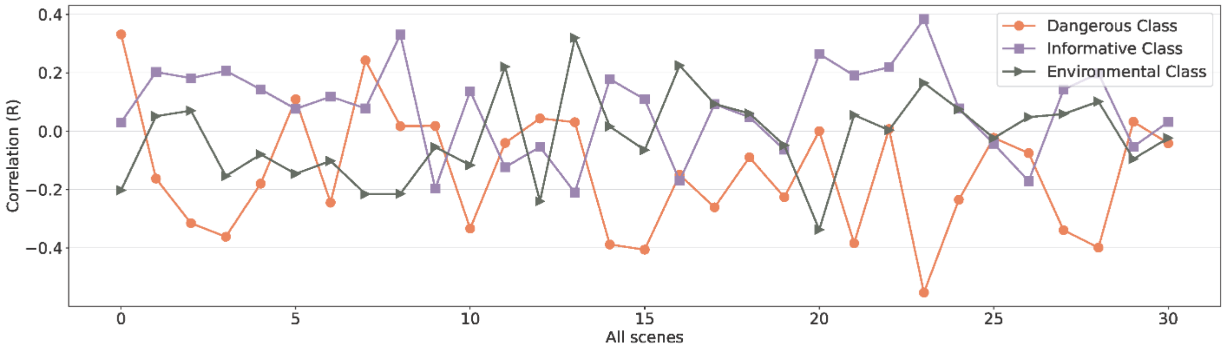

Figure 7. Linear correlation between object semantics and meaning

information category) are correlated with the meaning map and cannot be considered independent of their guidance of visual attention with meaning. Among them, the danger category is negatively correlated with the meaning map. Dangerous objects are more difficult to recognize than everyday scene objects because of the characteristics of breakage, loss of original order, confusion, and mismatch with inherent cognition, making semantics slightly negatively correlated in the representation of meaning. The distractors (environment class), on the other hand, are uncorrelated with the meaning map, which further suggests that distractors may be difficult to provide valuable recognition information in the task.

Since object semantics and meaning are partially correlated, the independent bootstrapping of salience, meaning, and semantics on attention was further investigated by multiple linear regression. Table 2 shows the independent influence of either factor on attention, controlling for the other factors. In particular, meaning independently explained 9.4% of the variation in attention (M = 0.094, s.d. = 0.092), whereas object semantics remained relevant to attention independently of significance, with each category explaining about 2% of the variation in attention. Relative to other factors, the guidance of attention by meaning remained dominant.

Table 2. Multiple linear regression result

Category	M(R^2)	S.D.	p-value
Meaning	0.094	0.092	5.3e-19
Saliency	0.050	0.055	5.9e-48
Danger	0.025	0.028	2.5e-4
Information	0.013	0.019	2.2e-12
Environment	0.018	0.024	2.5e-2

5 Conclusion and future works

Previous work has suggested numerous factors that will have an impact on visual attention, which will be directly expressed in human behavior and task efficiency and accuracy. Most of the past researches base their experiments on daily environment, using some real world or simplified feature expressions that people often come into contact with as stimuli, but in fact, high-stress and high-risk environments form stronger stimuli and cognitive loads, which make the human attention pattern more complex. At the same time, the analysis about scenes and semantics has mainly focused on the quantitative expression of a certain semantic feature, while failing to completely describe the post-disaster environment and the work faced by search and rescue personnel.

In this study, a post-earthquake cluttered environment was designed as an experimental stimulus, and routes and tasks were designed according to the remote search and rescue approach. Our methodological approach, employing eye-tracking technology alongside the development of scenario-specific feature maps, enabled an in-depth analysis of visual attention in high-stress SAR conditions. The interactions between the guiding factors of salience, meaning, task, and object were investigated, and it was found that the effect of any factor on attention was not independent and that meaning remained dominant. This suggests that the construction of an attention model needs to take multiple factors into account, and that a more comprehensive model will lead to a better explanation of attention. The inclusion of participants without SAR background revealed their capability to allocate attention to targets as per task demands, providing empirical support for mobilizing a broader workforce in post-disaster rescue efforts. This study also used semantic attribute maps to represent scene features and understand how objects affect visual attention. These findings provide a more comprehensive perspective for understanding human visual attention mechanisms.

In summary, the main findings of this study are:

- Meaning and visual salience are significantly correlated in post-disaster scenes. They both predicted the distribution of attention, but after controlling for the relationship between significance and salience, only significance contributed to the unique variance of the attention distribution.

- Objects do not direct attention independently of meaning in semantically related tasks, although the correlation between the two is weak.
- Object relative value is also able to direct visual attention, with guidance from the target object being more significant relative to guidance from distractors.

Although this study provides a robust analysis of the mechanisms of visual attention for teleoperated drivers in natural disaster scenarios, it does have some limitations. In the study of dynamic scenes, time and cognitive updating formed by repetitive images were not added to the analysis of attention to form a model of attention with a temporal sequence. As a result, the study did not provide a complete picture of the possible effects of successive visual images. Subsequent research could build on the foundation of this study and try to establish a more complete theory of visual attention and construct a more accurate predictive model of attention. Further work could also assist the operator to locate the target through multimodal modeling, give real-time positive guidance to attention, and improve the efficiency and success rate of search and rescue.

References

[1] S. Mandyam, S. Priya, S. Suresh, and K. Srinivasan, A Correlation Analysis and Visualization of Climate Change using Post-Disaster Heterogeneous Datasets, , 2022.

[2] United Nations Office for Disaster Risk Reduction Centre for Research on the Epidemiology of Disasters, The human cost of disasters: an overview of the last 20 years, , 2020.

[3] B. Mishra, D. Garg, P. Narang, and V. Mishra, Drone-surveillance for search and rescue in natural disaster, *Computer Communications*. 156:1–10, 2020.

[4] A.M. Treisman, The effect of irrelevant material on the efficiency of selective listening, *The American Journal of Psychology*. 77:533–546, 1964.

[5] D.E. Broadbent, Perception and communication, Elsevier, 2013.

[6] J.M. Wolfe, and T.S. Horowitz, Five factors that guide attention in visual search, *Nat Hum Behav*. 1:1–8, 2017.

[7] L. Itti, C. Koch, and E. Niebur, A model of saliency-based visual attention for rapid scene analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 20:1254–1259, 1998.

[8] J. Harel, C. Koch, and P. Perona, Graph-Based Visual Saliency, in: Advances in Neural Information Processing Systems, MIT Press, 2006.

[9] J.M. Henderson, and T.R. Hayes, Meaning-based guidance of attention in scenes as revealed by meaning maps, *Nat Hum Behav*. 1:743–747, 2017.

[10] J.M. Henderson, T.R. Hayes, G. Rehrig, and F. Ferreira, Meaning Guides Attention during Real-World Scene Description, *Sci Rep*. 8:13504, 2018.

[11] C.E. Peacock, T.R. Hayes, and J.M. Henderson, Meaning guides attention during scene viewing, even when it is irrelevant, *Atten Percept Psychophys*. 81:20–34, 2019.

[12] E.B. Foa, and M.J. Kozak, Emotional processing of fear: Exposure to corrective information, *Psychological Bulletin*. 99:20–35, 1986.

[13] J.D. Fraustino, J.Y. Lee, S.Y. Lee, and H. Ahn, Effects of 360° video on attitudes toward disaster communication: Mediating and moderating roles of spatial presence and prior disaster media involvement, *Public Relations Review*. 44:331–341, 2018.

[14] M.L.-H. Võ, The meaning and structure of scenes, *Vision Research*. 181:10–20, 2021.

[15] J. Fan, X. Li, and X. Su, Building Human Visual Attention Map for Construction Equipment Teleoperation, *Front Neurosci*. 16:895126, 2022.

[16] R. Eyraud, E. Zibetti, and T. Baccino, Allocation of visual attention while driving with simulated augmented reality, *Transportation Research Part F: Traffic Psychology and Behaviour*. 32:46–55, 2015.

[17] F. Driewer, K. Schilling, and H. Baier, Human-Computer Interaction in the PeLoTe rescue system, in: IEEE International Safety, Security and Rescue Rototics, Workshop, 2005., IEEE, 2005: pp. 224–229.

[18] D.J. Rea, S.H. Seo, N. Bruce, and J.E. Young, Movers, Shakers, and Those Who Stand Still: Visual Attention-grabbing Techniques in Robot Teleoperation, in: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, Association for Computing Machinery, New York, NY, USA, 2017: pp. 398–407.

[19] M.S. Castelhano, M.L. Mack, and J.M. Henderson, Viewing task influences eye movement control during active scene perception, *J. Vision*. 9:6, 2009.

[20] Y. Chen, and G.J. Zelinsky, Is There a Shape to the Attention Spotlight? Computing Saliency Over Proto-Objects Predicts Fixations During Scene Viewing, *J. Exp. Psychol.-Hum. Percept. Perform.* 45:139–154, 2019.