SELECTIVE VISUAL ATTENTION FOR URBAN SEARCH AND RESCUE (USAR) SYSTEMS

R. ABDULLAH, R. HAMILA, and D. VERNON

Kalifa University of Science, Technology, and Research
Sharjah, UAE
E-mail: rasha@kustar.ac.ae
www.kustar.ae.ac

In urban search and rescue mobile robotics, one of the most significant problems is to identify and localize trapped victims in collapsed or dangerous buildings. In order to optimize search operations, selective visual attention plays a crucial role, as it focuses on the visual stimuli that are most relevant when identifying victims, inhibiting other stimuli in the scene. This reduces the amount of resources required for subsequent visual processing, such as recognition, and typically results in more effective and quicker search strategies. In this work, we concentrate on a purely bottom-up computational model. It functions by differentiating high-contrast areas in the image using several features including intensity, colour, and visual motion. Results for artificial and real-world images are presented, showing that the implemented system has an acceptable performance in identifying moving human parts.

Keywords: Selective visual attention; Bottom-up; Saliency map; Feature maps.

1. Introduction

We investigate selective visual attention as an approach for the identification and localization of victims using semi-autonomous mobile robots involved in search and rescue operations in unknown environments. Selective visual attention is mainly used to reduce the amount of computation required to process irrelevant incoming visual information, in order to maximize relevant information thus making it more suitable for further processing and/or for visual interpretation[1, 2, 3]. Consequently it reduces the search time required to estimate the victim's location. It employs a method of determining the most salient visual stimuli in the visual scene, selecting it, and suppressing other stimuli[3, 4, 5].

The saliency of a location can be determined by either bottom-up attention or top-down attention[1, 2, 6]. Bottom-up attention can be determined

based on the contrast in visual features of the target and its surrounds in visual scenes[3, 6, 7, 8]. It is mainly determined by a visual search mechanism. The visual search looks for a target in the visual scene based on one distinctive visual property, known as pop-out search, or a group of distinctive properties, known as conjunctive search[5, 9, 10, 11]. In contrast, top-down attention can be determined by the prior-knowledge of the target or its environment[1, 3, 6, 7, 8].

In order to deploy selective attention in the real world, we built the attentive system based on the Itti *et al.* computational model[12]. It is a bottom-up computational model, that concentrates on the bottom-up information. It can be broken up into: pre-attentive computations and attentive computations[2, 4]. Pre-attentive computations consist in forming an abstract representation of the raw incoming visual information. On the other hand, attentive computations consist in using an abstract representation generated by pre-attentive computations to determine the location to which the attention should be drawn[4, 11, 13].

Pre-attentive computations can be further broken up into extraction visual features from input images, combination of these features to produce a saliency map, initially proposed by Koch and Ullman[13]. It is a two-dimensional map that presents the salience of each visual stimulus in the visual scene[1, 2, 4, 14].

Attentive computations can be further broken up into control strategies: a Winner Take All (WTA) and an Inhibition Of Return (IOR). The first strategy identifies a point of the highest salience value. The second strategy identifies a point of the next highest salience value after temporarily inhibition of the highest salience value[1, 2, 4, 12, 14].

In this paper, we present the architecture of the implemented computational model in Sec. 2. Results of artificial and natural experiments are discussed in Sec. 3. Finally, Sec. 4 summarizes the presented work and future directions.

2. A Framework of Selective Visual Attention

The Itti *et al.* model is a purely bottom up computational model. It extracts pre-attentive modalities of intensity, colour (Red, Green, Blue, and Yellow), and orientation (0, 45, 90, and 135). These modalities are assembled into a multi-scale representation using Gaussian pyramids[15]. Within each modality, a center-surround mechanism is applied in order to generate multi-scale feature maps using Laplacian pyramids[15]. A multi-scale combination is then employed to transform these multi-scale feature maps into

conspicuity maps, which represent the saliency of each modality. Finally, conspicuity maps are linearly combined to determine the saliency of each location of the visual scene. In order to combine multi-scale maps, maps are normalized based on the contents-based global no-linear amplification strategy, that functions on the difference between the global maximum and the average of local maxima in each feature map and conspicuity map [6, 12, 16].

The original Itti model is intended to provide the most salient location on static images, and therefore omits a component of motion. However, motion can be an important feature to identify the target in dynamic scenes. Consequently, we extract the magnitude of motion velocity of each point instead of orientation. This magnitude is extracted by computing the number of moving pixels of each point between image frames f_0 and f_1 using an instantaneous optical flow motion segmentation technique[17]. We also normalize multi-scale maps using the linear learning strategy[6, 16], in addition to the contents-based global amplification strategy. In this strategy, weights of each feature type is learned based on the comparison of the map's response inside and outside the desired target, which is skin-colour in this case.

The original Itti model is also a biologically-inspired model. It uses a Winner Take All (WTA) neural network to find the most salient point in the saliency map. We use a seeded region growing segmentation technique, proposed by Frintrop *et al.*[11, 18] instead of using WTA networks. Despite it is less biologically plausible, it yields equivalent results and requires less computations[11, 18]. The computation of the most salient point is determined by finding the maximum value in the saliency map and accepting all surround values that differ at most 0.80 from the maximum value. The implemented architecture is shown in Fig. 1.

3. Experiments and Results

To illustrate the performance of the implemented model, we ran a number of experiments based on synthetic and real-world images.

3.1. *Search Performance in Synthetic Images*

We have tested the model with different synthetic images to establish the link to the human visual attention[6, 8, 11]. These simplified images are allowed us to examine the functioning of each component of the model.

We designed two sets of synthetic images: pop-out test images and conjunctive test images. In pop-out test images, the target can only be distin-

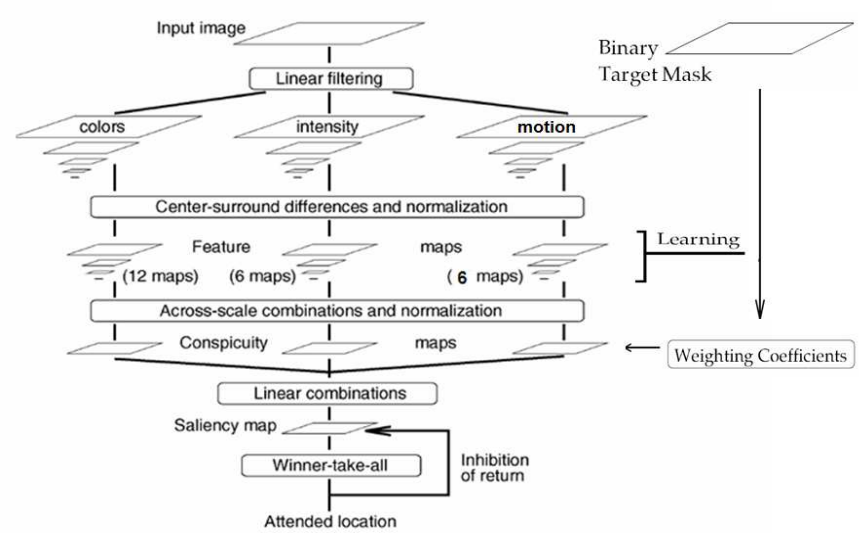

Fig. 1. A framework of the implemented system

guished by one visual feature *e.g.* intensity, colour, or motion. However, in conjunctive test images, the target can be distinguished by more than one visual feature *e.g.* intensity and motion conjunction or colour and motion conjunction. In this paper, we will discuss the neighbourhood influence on the target's salience in pop-out and conjunctive images.

In the first experiment, the number of homogeneous neighbours, which have at least one feature that is different to the target, is increased. The goal of this experiment is to show that when the number of such homogeneous neighbours increases, the target's salience increases.

In pop-out images, green vertical bars are gradually added in the neighbourhood of the red vertical target. However, they are added to the moving red target in conjunctive images.

The results of pop-out images show increasing target's salience with increasing the number of homogeneous neighbours. The results of conjunctive images show increasing target's salience with increasing the number of different visual features comparing with pop-out results. In addition, the target's salience increases with increasing the number of non-targets, see Fig. 2.

In the second experiment, the number of heterogeneous neighbours, which share at least one visual feature with the target, is also increased.

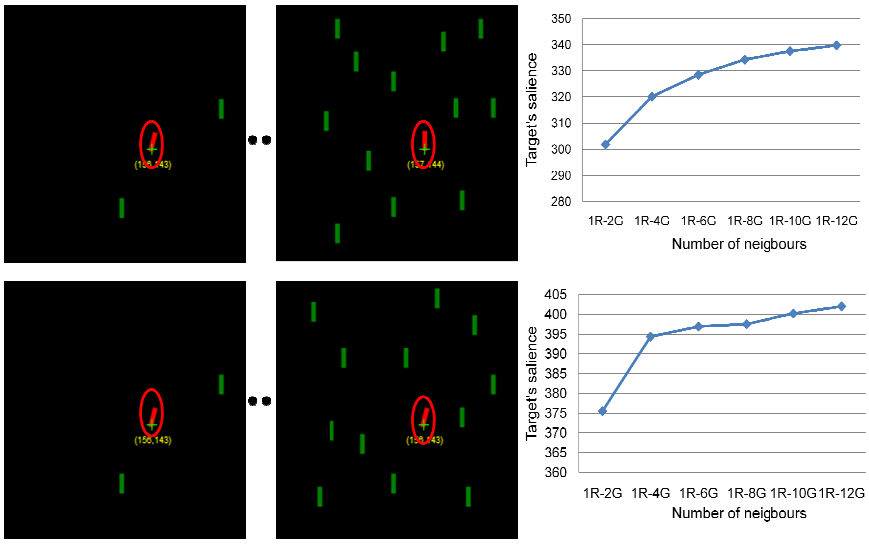

Fig. 2. The scaling effect of uniform neighbours: the first row presents results of pop-out images, and the second row presents results of conjunctive images.

The goal of this experiment is to show that when the number of such heterogeneous neighbours increases, the target's salience decreases.

In the first set of conjunctive images, red vertical bars are gradually added in the neighbourhood of the moving red target, since it is rotated by 15 degree. On the other hand, added red bars are rotated by a lesser degree in the second set of conjunctive images.

The results show that the targets salience becomes weaker as more neighbours share the same colour as the target in the first set, but stronger than in the second set. The reason is that in the first set, the strength is a result of the contrast among the moving red target, vertical red bars, vertical green bars and its background. In the second set, the contrast between moving red target and other moving red bars decreases. In other words, stronger contrast between the target and its neighbourhood makes the target more salient to capture visual attention in the bottom-up competition, see Fig. 3.

3.2. *Search Performance in natural scenes*

We generated six experimental data sets to evaluate the system behaviour in natural images. In the first set, real-world images contain a non-moving

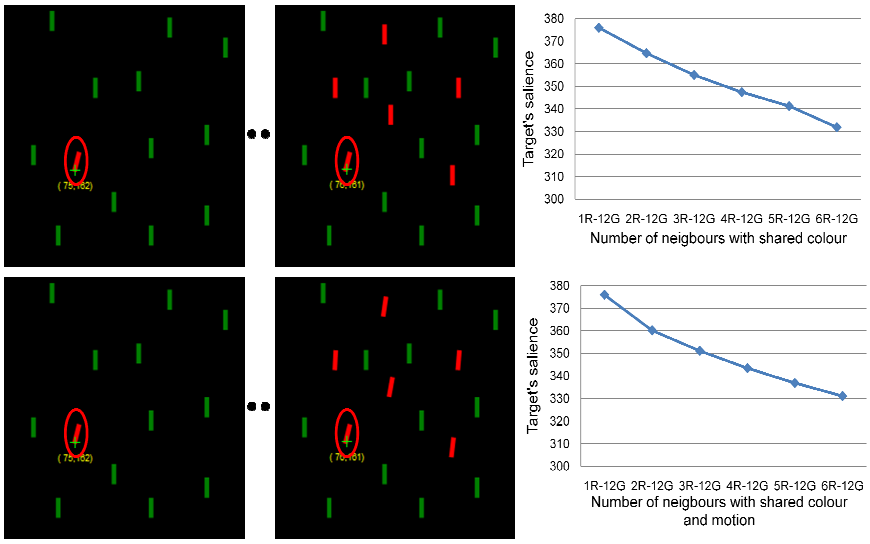

Fig. 3. The similarity effect of neighbours: the first row presents results of the first set of conjunctive images, and the second row presents results of the second set of conjunctive images.

hand in a clear background. In the second set, images also contain a non-moving hand, but in a cluttered background. Real-world images of third and fourth sets, have a moving hand, but images in the fourth set are cluttered backgrounds. Colour images of fifth and sixth sets do not contain a hand. In the fifth set, images contain only one object; however, images of the sixth set contain many objects.

The results in the first set show that the model does not reliably attend to the human part. The reason is that the static part of the model detects illumination variations, and colour variations, especially Red/Green contrast and Blue/Yellow contrast. On the other hand, the dynamic part of the model detects motion variations, so it detects the moving object in dynamic environments. Consequently, the FOA is directed based on illumination, colour, or/and the magnitude of the movement velocity. However, in this set, we omit the motion component, so the FOA is directed based on intensity and colour variations only. In the second set, the implemented model fails to attend to human parts, because of higher intensity and colour variations.

In the third set, the implemented model can distinguish between moving and non-moving objects, since the motion conspicuity map exhibits a high

activation (bright spot) at regions of the desired property, highest magnitude of the motion velocity, and lower activation (darker) at other regions. However, the results of the fourth set is less accurate than the third set, since both intensities and colours of different objects in real-world images affect on determining the most salient point in the saliency map. However, the performance of the implemented system can be improved by combination a skin segmentation map with other conspicuity maps. An example of the search performance in the fourth set is shown in Fig. 4.

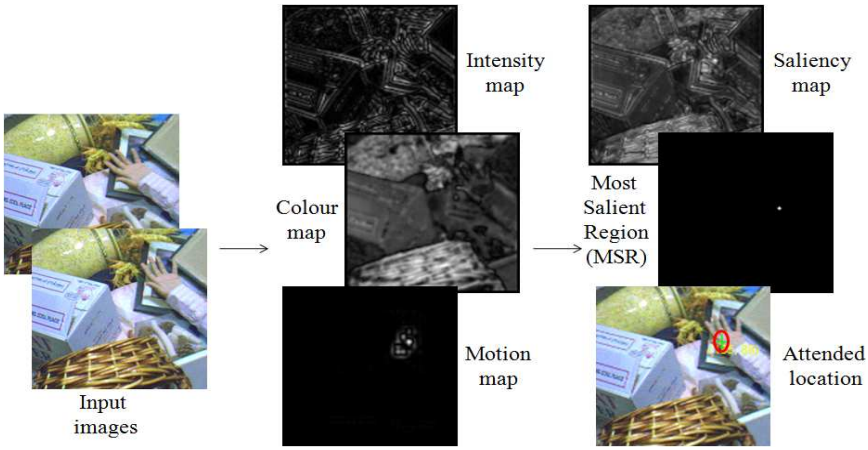

Fig. 4. An example of the search performance in cluttered scenes

In the fifth set, the model identifies the most salient point as a result of uniqueness intensity and colour. On the other hand, the saliency variation in the sixth set is larger than in the fifth set, therefore results are worse.

There are other approaches to evaluate the performance of the implemented model in synthetic and natural images could not mentioned in this paper.

4. Conclusion and Future Work

We have presented a selective attention system to find victims for semi-autonomous robots in rescue scenarios. Although the attentive system yields an acceptable performance, there are a lot could be done to improve the performance of the system. The first issue to be discussed is extending the attentive system to include high-level visual processes, such as human

recognition since it helps to reduce the required time to attain to human parts. Another issue to be further investigated is extending the attention to different perceptual features, not only visual features, such as an integration of visual and audio attentions.

References

1. A. L. Rothenstein and J. K. Tsotsos, *Image Vision Computing* **26**, 114 (2005).
2. L. Itti, "Models of Bottom-Up Attention and Saliency", in *Neurobiology of Attention*, (Elsevier, 2005) pp. 576–582.
3. G. Backer, B. Mertsching and M. Bollmann, *IEEE Transactions Pattern Analysis Machine Intelligence* **23**, 1415 (2001).
4. L. Itti and C. Koch, *Nature Reviews Neuroscience* **2**, 194 (2001).
5. L. Itti, "visual attention", in *The Handbook of Brain Theory and Neural Networks*, ed. M. A. Arbib (MIT Press, 2003) pp. 1196–1201.
6. L. Itti, *Models of Bottom-Up and Top-Down Visual Attention, Ph.D Thesis* (California Institute of Technology, 2000).
7. K. Papantzikos and N. Tsapatsoulis, "On the implementation of Visual Attention Architecture", in *Tales of the Disappearing Computer*, 2003.
8. Y. Sun and R. Fisher, *Artifical Intelligence* **146**, 77 (2003).
9. L. Itti and C. Koch, *Vision Research* **40**, 1489 (2000).
10. E. Niebur, L. Itti and C. Koch, "controlling the focus of visual selective attention", in *Models of Neural Networks IV*, ed. L. Van Hemmen and E. Domany and J. Cowan (Springer Verlag, 2001)
11. S. Frintrop, *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search, PH.D Thesis* (Rheinische Friedrich-Wilhelms-Universitat Bonn Institut Fur Informatik and Freunhofer Institut fur Autonome Intelligente Systeme, 2006).
12. L. Itti, C. Koch and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (11)1998.
13. C. Koch and S. Ullman, *Human Neurobiology* **4**, 219 (1985).
14. D. Walther, *Interactions of Visual Attention and Object Recognition: Computational Modeling, Algorithms, and Psychophysics, PHD Thesis* (California Institute of Technology Pasadena, California, 2006).
15. P. Burt and E. Adelson, "The Laplacian Pyramid as a Compact Image Code", in *IEEE Transactions on Communications*, (4) (Signal Processing and Communication Electronics of the IEEE Communications Society, 1983).
16. L. Itti and C. Koch, *Journal of Electronic Imaging* **10**, 161 (2001).
17. D. Vernon, *Image and Vision Computing* **17**, 189 (1999).
18. S. Frintrop, P. Jensfelt and H. Christensen, "Simultaneous Robot Localization and Mapping Based on a Visual Attention System", in *Attention in Cognitive Systems, Lecture Notes on Artificial Intelligence (LNAI)*, (Springer-Verlag, 2007).