A Fast Visual Search and Recognition Mechanism for Real-Time Robotics Applications

Quoc Vong Do[1], Peter Lozo[2], and Lakhmi Jain[3]

[1,3] Knowledge-Based Intelligent Engineering Systems Center, University of South Australia, Mawson Lakes, S.A. 5095, Australia
Quoc.Do@postgrads.unisa.edu.au, Lakhmi.Jain@unisa.edu.au
[2] Weapons Systems Division, Defence Science and Technology Organisation, PO Box 1500, Edinburgh, SA 5111
peter.lozo@dsto.defence.gov.au

Abstract. Robot navigation relies on a robust and real-time visual perception system to understand the surrounding environment. This paper describes a fast visual landmark search and recognition mechanism for real-time robotics applications. The mechanism models two stages of visual perception named pre-attentive and attentive stages. The pre-attentive stage provides a global guided search by identifying regions of interest, which is followed by the attentive stage for landmark recognition. The results show the mechanism validity and applicability to autonomous robot applications.

1 Introduction

Autonomous robot navigation needs a reliable and robust visual perception system to gain an understanding and awareness of the surrounding environment. Such a system could be modeled based on the effectiveness and robustness of a human visual system. Observation of the human visual system indicates that people are capable of quickly detecting a flying object and successfully avoiding collision, without the needs of object recognition. The identification of the flying object comes after the collision avoidance behaviour. This leads to a suggestion that human visual perception has two stages named pre-attentive and attentive [1]. Pre-attentive stage is a fast global process, which aims at identifying regions of interest (ROI) that are most likely to have the target object embedded within it. In comparison the attentive stage is a high level process that identifies the objects within the selected ROI regions. This is a slow computationally intensive process.

In general, when encountering a visual scene, people tend to focus on the most 'eye catching', contrasted or coloured regions. This ability can be modeled by the pre-attentive stage, where the input image is quickly analysed to determine ROI regions prior to any thorough object search. This allows the system to concentrate on ROI regions and provide a guided search mechanism for fast object recognition in the attentive stage.

Many attempts to model the pre-attentive process of visual perception have been reported in recent literature. The common methods are to detect the most 'stand out'

G.I. Webb and Xinghuo Yu (Eds.): AI 2004, LNAI 3339, pp. 937–942, 2004.

regions based on color, features and high contrast. In [2, 3], the ROI regions are selected using a colour segmentation, prior to landmark recognition using genetic algorithms. In [4], the most highly contrastive regions are considered as ROI regions, where landmark recognition is performed using selective attention adaptive resonance theory (SAART) neural networks, starting from the highest contrastive region to the lowest region. ROI regions are larger than memory images such that a thorough landmark search and recognition is performed within each selected ROI region. This paper presents an alternative implementation of the pre-attentive stage by using knowledge from memory images to select ROI regions for landmark recognition.

2 Pre-attentive and Attentive Stages for Visual Landmark Search and Recognition

The proposed visual landmark search and recognition architecture mimics the concepts of pre-attentive and attentive stages in the human vision system for landmark recognition. Although the proposed architecture is a simpler system, with fewer functions and may be subjected to minor violations to the actual human vision system, the architecture is capable of providing fast and robust landmark search and recognition. The overall architecture shown in Figure 1 is divided into two distinct, bottom and top sections, to model the pre-attentive and attentive stages of visual perception respectively. Initially, a grey level input image of 240x320 pixels resolution is pre-processed using a 3x3-mask Sobel edge detection. The edge image is used to produce a dilated image using a 5x5 averaging window, where each pixel in the dilated edge image summons the average edge activities over a local 5x5-region in the Sobel edge image. This process is used to achieve distortion invariant landmark recognition [5], where the distorted edges are compensated by adjacent neighboring pixels. The dilated image is then passed through the pre-attentive stage, which involves the determination of ROI regions and further processes each ROI region to classify as potential regions (PR). Only PR regions are passed into the attentive stage for landmark recognition, all remaining regions are discarded.

The determination of PR regions uses the knowledge obtained from memory images to calculate two thresholds for each landmark: ROI and signature thresholds. First of all, considers three memory images of three different landmarks shown in Figure 2, a total number of significant pixels which describes a shape of each landmark is calculated by comparing each pixel against a small threshold to remove weak edges. A ROI threshold is set for each landmark, which is equal to 50% of the total number of significant pixels of the corresponding memory image. Signature thresholds on the other hand are calculated based on edge activities of internal features of each landmark, named unique region(s). These regions are fixed regions, describing physical internal appearances of each landmark and are unchanged from the camera field of views. The signature threshold is set to be equal to the number of significant pixels in the selected unique region(s) of each landmark.

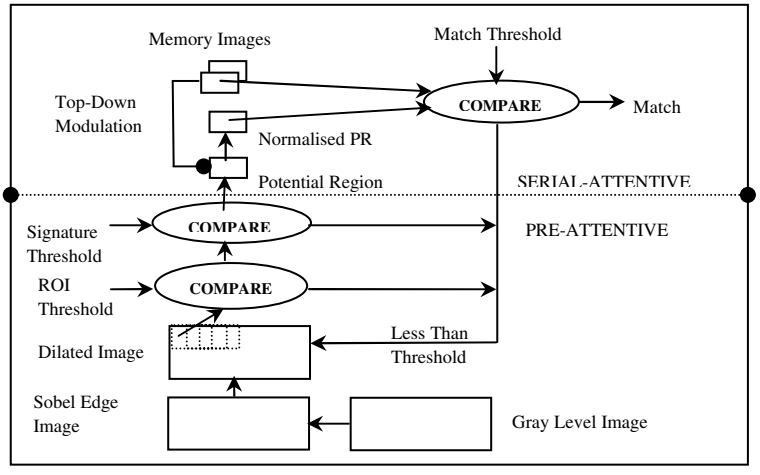

Fig. 1. The overall image processing architecture that combines the pre-attentive and the attentive stages

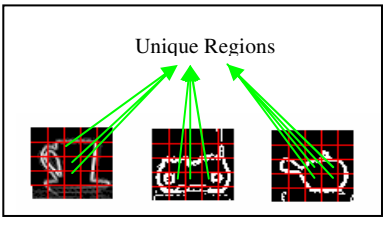

Fig. 2. Three edge detected images of selected landmarks. Regions indicated are unique regions, used for the determination of the signature threshold for each landmark

The determination of PR regions and ROI regions are based on the comparison of input regions (within the 50x50 search window) against both the ROI and signature thresholds. The input region that is greater than ROI threshold is classified as the ROI region. Then each ROI region is subjected to a further comparison with the signature threshold before being promoted into the attentive stage, where it is subjected to intensive image processing for landmark recognition.

In the attentive stage, the landmark recognition architecture is developed based on previous works [4-7], where a vision-inspired SAART neural network is proposed for landmark recognition. The SAART neural network is a derivation of adaptive resonance theories (ART) proposed by Grossberg and Carpenter [8, 9]. It incorporates an additional mechanism for top-down memory selective attention, which is achieved by pre-synaptic modulation of the input to facilitate relevant portions and inhibit irrelevant portions of input images. Thus enables the system to recognise known objects embedded in clustered images. The SAART neural network is a dynamic system and thus computationally intensive. Therefore, instead of using the whole network, the

developed architecture uses only the top-down memory modulation mechanism. This mechanism uses the knowledge from memory to assist the early stage of features extraction, which increases the robustness of the landmark recognition system. Each extracted region is subjected to template matching with the corresponding memory image using the cosine rule between two 2-D images. This results in a match value range from 0-1 (where 1 represents 100% match), which is evaluated against a match threshold of 90%.

3 Results and Discussions

The performance of the developed pre-attentive stage is evaluated by measuring the time taken to completely process a series of selected scenes both with and without the pre-attentive stage. Five different scenes of different indoor environment are selected to demonstrate the effectiveness of the pre-attentive stage. The first input scene is selected with the landmark embedded in a clean background to provide an insight into the system ideal performance. The scene-2 and sense-3 are selected from office environment, and sence-4 and sence-5 are selected from a corridor environment. Figure 3 shows different image processing stages for scene-4. Initially, the system performed Sobel edge-detection on an input grey level image producing an edge image, which is blurred using a 5x5 averaging mask as shown in Figure 3(b) and Figure 3(c) respectively. The blurred image is than entered the pre-attentive stage, where a 50x50 search window is scanning across the image for PR regions determination, which is followed by landmark recognitions in the attentive stage. The results of the landmark search and recognition process are converted into a range from 0-255 and displayed as a grey level image in Figure 3(d), with the highest level of contrast represents the highest match value. The black regions are ones that have been skipped by the pre-attentive stage. The landmark is found at a location, where the match value is greater than the match threshold.

Table 1. The time taken to process each selected scene

Algorithms	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5
Without Pre-Attentive	6.91s	7.12	7.02s	7.098s	6.987s
Pre-Attentive ROI Threshold	0.423s	4.051s	4.092s	3.246s	2.711s
Pre-Attentive ROI Threshold & Signature Threshold	0.276s	2.206s	3.665s	2.926s	1.328s

The time taken to process each selected scene is summarised in Table 1. The system takes 7.025s on average to process the input image without the pre-attentive stage. The system performance has improved significantly with the integration of the

pre-attentive stage. For scene-1 the system is able to reduce the processing time to 0.423s using the ROI threshold and with a further reduction to 0.276s using signature threshold. This is the system ideal performance in clean background. In the office and corridor environments, scene-2 to scene-5, the processing time is reduced to approximately 2-4 seconds, with a further reduction to 1 to 2 seconds by applying the signature threshold.

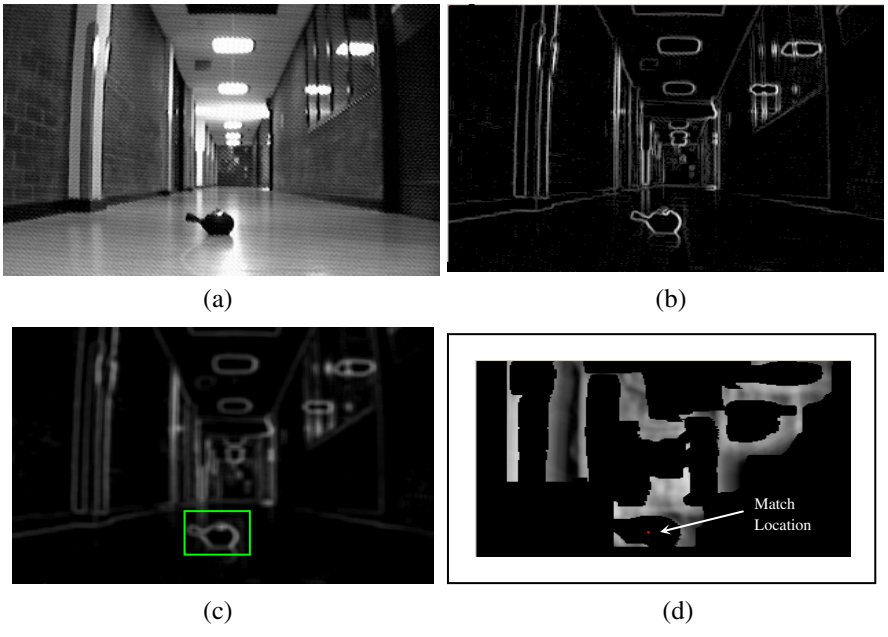

Fig. 3. A sample processed frame along a corridor. (a) Gray level input image, (b) Sobel edge detected image, (c) Dilated edge image, (d) Degree of match between memory and each region in the input scene, which is converted into a range from 0-255 and displayed as a grey level image

4 Conclusions

This paper has presented a fast visual search and recognition image processing architecture for real-time robotics applications. The architecture incorporates a simple implementation of pre-attentive and attentive stages for fast and robust visual landmark search and recognition. The proposed pre-attentive stage is able to reduce the recognition time from seven seconds to approximately 0.276 second depending on the amount of edge activities in the visual scene. The improvement in landmark recognition speed provides for real-time applications to autonomously navigating robots.

However, further developments to this work are required to cope with various robot navigation speeds. As the robot navigates, the size of the landmark changes con-

stantly. This requires the robot to be capable of size invariant landmark recognition. Similarly, the appearance of the landmark depends on the approaching angle, which leads to a requirement for 2D aspect view invariant landmark recognition.

Acknowledgment

The work described in this paper was funded by Weapons Systems Division of DSTO via research contract No. 4500 177 390.

References

1. B. Juesz and J. R. Bergen, "Texons, the Fundamental elements in pre-attentive vision and perception of textures," *Bell System Technical Journal*, vol. 62, pp. 1619-1645, 1983.
2. M. Mata, J. M. Armingol, A. de la Escalera, and M. A. Salichs, "A visual landmark recognition system for topological navigation of mobile robots," presented at The IEEE International Conference on Robotics and Automation, Proceedings 2001 ICRA., pp.1124-1129, 2001.
3. M. Mata, J. M. Armingol, A. de la Escalera, and M. A. Salichs, "Using learned visual landmarks for intelligent topological navigation of mobile robots," presented at IEEE International Conference on Robotics and Automation, Proceedings. ICRA-03, pp.1324-1329, 2003.
4. E. W. Chong, C.-C. Lim, and P. Lozo, "Neural model of visual selective attention for automatic translation invariant object recognition in cluttered images," presented at Knowledge-Based Intelligent Information Engineering Systems, 1999. Third International Conference, pp.373-376, 1999.
5. J. Westmacott, P. Lozo, and L. Jain, "Distortion invariant selective attention adaptive resonance theory neural network," presented at Third International Conference on Knowledge-Based Intelligent Information Engineering Systems, USA, pp.13-16, 1999.
6. P. Lozo and C.-C. Lim, "Neural circuit for object recognition in complex and cluttered visual images," presented at The Australian and New Zealand Conference on Intelligent Information Systems, pp.254-257, 1996.
7. P. Lozo, "Neural Circuit For Self-regulated Attentional Learning In Selective Attention Adaptive Resonance Theory (saart) Neural Networks," presented at The Fourth International Symposium on Signal Processing and Its Applications, ISSPA-96, pp.545-548, 1996.
8. S. Grossberg and L. Wyse, "Invariant recognition of cluttered scenes by a self-organizing ART architecture: figure-ground separation," presented at International Joint Conference on Neural Networks, IJCNN-91-Seattle, pp.633-638, 1991.
9. G. A. Carpenter, S. Grossberg, and D. Rosen, "ART 2-A: an adaptive resonance algorithm for rapid category learning and recognition," presented at International Joint Conference on Neural Networks, IJCNN-91-Seattle, pp.151-156, 1991.