Proceedings Paper:

Jaramillo-Avila, U., Hartwell, A., Gurney, K. orcid.org/0000-0003-4771-728X et al. (1 more author) (2018) Top-down bottom-up visual saliency for mobile robots using deep neural networks and task-independent feature maps. In: Giuliani, M., Assaf, T. and Giannaccini, M., (eds.) Towards Autonomous Robotic Systems. TAROS 2018, 25-27 Jul 2018, Bristol, UK. Lecture Notes in Computer Science, 10965 . Springer Verlag , pp. 489-490. ISBN 9783319967271

https://doi.org/10.1007/978-3-319-96728-8

The final publication is available at Springer via https://doi.org/10.1007/978-3-319-96728-8

Top-down Bottom-up Visual Saliency for Mobile Robots using Deep Neural Networks and Task-Independent Feature Maps

Uziel Jaramillo-Avila[1], Adam Hartwell[1], Kevin Gurney[2], Sean Anderson[1]

1. Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S1 3JD, UK
2. Department of Psychology, University of Sheffield, Sheffield, S10 2TN, UK

{ujaramilloavila1, ahartwell1, k.gurney, s.anderson}@sheffield.ac.uk

Visual saliency is a biological mechanism for shifting visual attention to important objects in the environment, where important objects could be hazards, or items associated with a task [1]. This approach to analysing visual scenes reduces the computational burden on vision systems by only focusing on a few important stimuli rather than the whole scene. Visual saliency is therefore potentially important for robots, to enable effective and safe operation in unstructured environments [2].

Visual saliency models can contain a bottom-up and/or a top-down component [3]. The bottom-up and top-down components have particular, respective advantages. The bottom-up component is typically fast and does not require training via machine learning algorithms. Bottom-up methods should also be more robust, because they do not require object recognition to operate successfully. The top-down component is essential, however, for task-dependent actions where a robot would need to recognise important objects to complete a task.

Machine learning algorithms are typically used for top-down visual saliency, e.g. via support vector machines [4]. Recently, deep neural networks have been applied to the task of top-down bottom-up visual saliency, exploiting the accuracy of deep networks in image recognition for the top-down component [5]. However, methods that combine deep neural networks for top-down visual saliency and task-independent feature maps for bottom-up saliency have not yet been developed. This is a gap in the literature, which this work aims to address.

In this investigation, we combined a well-known approach to bottom-up visual saliency, using task-independent feature maps, based on e.g. colour contrast, intensity contrast and orientation contrast [6], with a deep convolutional neural network (CNN) based on the Tiny-YOLO architecture [7]. The resulting bottom-up and top-down saliency maps (spatially calibrated maps of salient features) were fused using a weighted sum, which combined the output of both processing streams into a single saliency map to drive visual attention [8].

To evaluate the visual saliency scheme we generated data from a small mobile robot, a Turtlebot, in an indoor environment, and implemented the visual saliency algorithm on an NVIDIA Jetson TX2, which is a processing board for embedded systems with a small GPU (256 CUDA cores). The Jetson TX2 was also used to control movements of the Turtlebot, using an installation of Robot

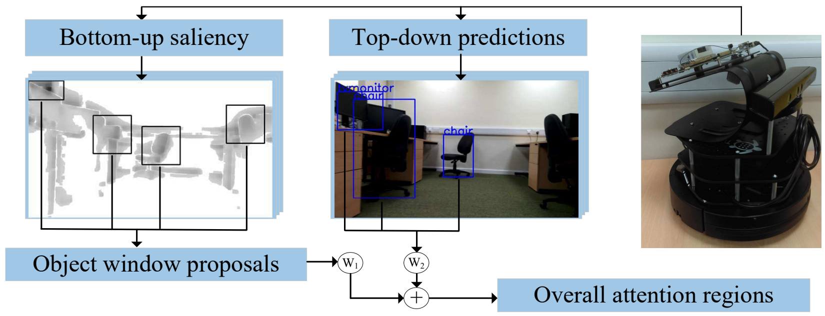

Fig. 1. Scheme for top-down bottom-up visual saliency. The Turtlebot with mounted Jetson TX2, on the right, generates and processes each map. In this example the top-down CNN fails to detect the chair to the right of the image, which is successfully detected by the task-independent map in the bottom-up pathway.

Operating System (ROS), so the control and visual processing was integrated on a single board.

We found that the bottom-up and top-down components worked as expected. However, a key early result from this pilot study was the observation that the top-down CNN would occasionally miss objects in the environment that were more robustly detected by the bottom-up task-independent feature maps (Fig. 1). Therefore, the fusion of bottom-up task-independent maps and top-down deep net maps appears promising for robust visual saliency in mobile robots.

References

1. Itti, L. and Koch, C.: Computational modelling of visual attention. Nature Reviews Neuroscience, 2(3), p.194–203 (2001)
2. Jiang, L., Koch, A. and Zell, A.: Salient regions detection for indoor robots using RGB-D data. IEEE International Conference on Robotics and Automation, 1323–1328 (2015)
3. Borji, A. and Itti, L.: State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1), 185–207 (2013)
4. Judd, T., Ehinger, K., Durand, F. and Torralba, A.: Learning to predict where humans look. IEEE International Conference on Computer Vision, 2106–2113 (2009)
5. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and VQA. arXiv Preprint, arXiv:1707.07998 (2017)
6. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence, 20, 1254–1259 (1998).
7. Redmon, J., Farhadi, A.: YOLO9000: Better, Faster, Stronger. IEEE International Conference on Computer Vision and Pattern Recognition, 6517–6525 (2017)
8. Kimura, A., Yonetani, R. and Hirayama, T.: Computational models of human visual attention and their implementations: A survey. IEICE Trans. on Information and Systems, 96(3), 562–578 (2013)