Biological Models for Active Vision: Towards a Unified Architecture

Kasim Terzić, David Lobato, Mário Saleiro, Jaime Martins, Miguel Farrajota,
J.M.F. Rodrigues, and J.M.H. du Buf

Vision Laboratory (LARSyS), University of the Algarve, Faro, Portugal
{kterzic,dlobato,masaleiro,jamartins,mafarrajota,jrodrig,dubuf}@ualg.pt

Abstract. Building a general-purpose, real-time active vision system completely based on biological models is a great challenge. We apply a number of biologically plausible algorithms which address different aspects of vision, such as edge and keypoint detection, feature extraction, optical flow and disparity, shape detection, object recognition and scene modelling into a complete system. We present some of the experiments from our ongoing work, where our system leverages a combination of algorithms to solve complex tasks.

1 Introduction

The problem of understanding complex visual scenes has been tackled from two main directions: computational approaches from computer vision, and the study and imitation of biological vision systems. Scene understanding systems attempt to provide the best explanation of the observed scene in terms of a semantic, meaningful description of low-level image data, typically by describing scene objects and relations between them. They combine different vision algorithms and use top-down and bottom-up processing in order to solve what is known to be an NP-complete problem [1].

It is known that primate brains solve this problem with apparent ease, so the study of biological vision has played an important role since the beginnings of computer vision and the insights from neurological observations have resulted in many biologically inspired algorithms addressing sub-problems of scene understanding, primarily in the fields of object recognition and robotics. However, to our knowledge there is no vision system combining many different aspects of vision into an integrated and comprehensive biologically plausible system for active real-time vision. In this paper, we present our work towards such a system and provide examples of our system solving a number of different vision problems. We concentrate on the system architecture and algorithms working in combination. More detailed descriptions of individual methods can be found in our previous publications.

2 Related Work

There is a wealth of scene understanding systems roughly divided into four major streams: grammars [2–4], blackboard architectures [5, 6], probabilistic models

M. Chen, B. Leibe, and B. Neumann (Eds.): ICVS 2013, LNCS 7963, pp. 113–122, 2013.
© Springer-Verlag Berlin Heidelberg 2013

[7–9], and artificial intelligence methods based on ontologies and description logics [10–12]. Some systems perform active vision tasks by controlling cameras [13].

Many biologically-inspired algorithms for solving sub-tasks of the complete vision problem have been proposed, particularly for feature extraction and early vision [14, 15], attention [16], and object recognition, with hierarchies based on Gabor responses [17], convolutional nets [18], and a number of connectionist architectures [19–21]. Recently, complete plausible models of the ventral (object recognition) pathway have appeared, based on the HMAX model [22]. Modern robotics has also embraced biological algorithms, with biologically-inspired SLAM algorithms [23, 24], obstacle avoidance, and complete robotic architectures [25]. Attempts to build a comprehensive biologically plausible system have focussed on dynamic field-based models of different aspects of vision [26] and cognitive robots [25]. A good comparative summary of computer vision and biological vision is given in [27].

3 System Overview

Figure 1 gives an overview of our system, which is a simplified model of the mammalian brain (for an excellent overview of different visual processing pathways we refer to [28]). All modules share information in the form of maps of neural activations (population codes) which excite or inhibit neuron populations within each module. In the rest of this section, we briefly describe individual modules of our system. For more detailed information on individual algorithms, we refer to our previous publications.

3.1 Early Vision

Early vision refers to cortical areas V1 and V2, which perform low-level processing and provide input for both the ventral and dorsal pathways. We do not yet

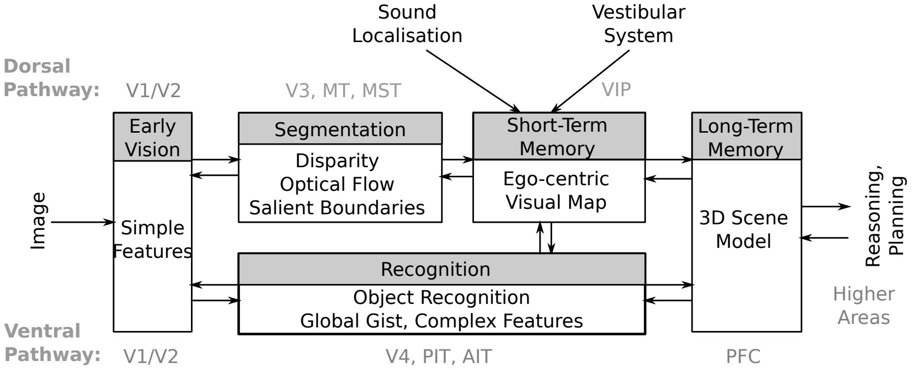

Fig. 1. Overview of our biologically-inspired active vision system. The top path models the dorsal pathway (localisation, motion and attention), bottom path the ventral pathway (recognition). Grey text indicates corresponding cortical areas.

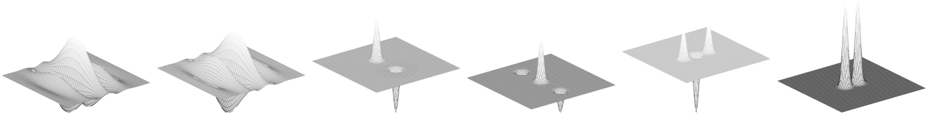

Fig. 2. Filter kernels in V1. From left to right: even simple cell, odd simple cell, single-stopped cell, double-stopped cell, tangential inhibition cell, radial inhibition cell.

model earlier processing in the retina and the lateral geniculate nucleus, notably the non-standard retinal ganglion cells.

For our V1 model, we use a multi-scale adaptation of Heitger's work [29]. Simple cells are modelled as complex Gabor filters, with a cosine-based real part, and a sine-based imaginary part. Complex cells are modelled by the magnitude of the simple cell responses, while the phase contains important information for disparity processing. We apply simple and complex cells at multiple scales (different Gabor wavelengths) and eight orientations. Complex cells are the basis for line/edge and keypoint detection. Line and edge detection is performed at locations where the complex cell response is maximum along the filter orientation. If at such a location one of the two components of the simple cell filter contains a zero crossing and the other one is maximum or minimum, an event is detected. There are four possible combinations of zero crossing and extrema, corresponding to four types of events: positive line, negative line, positive edge and negative edge. For each detected event, we keep the event type and the orientation of the strongest complex cell response.

For keypoint extraction, we use models of end-stopped cells. Single-stopped and double-stopped cells are modelled as a mixture of Gaussians, which respond to line/edge terminations and corner/blob-like features. We apply two types of inhibition to suppress responses along lines and edges. An overview of the cell models is given in Fig. 2, and example results in Fig. 3. For a detailed mathematical model, we refer to [29]. At V2 level, we extract curve segments based on Gestalt principles of good continuity, extract symmetries and group low-level events into simple descriptors.

We have three implementations of our V1 model. The CPU version is competitive with computational interest point detectors like SIFT and SURF, while the two GPU versions (based on CUDA and OpenCL) easily run in real time.

3.2 Segmentation and Attention

The dorsal "where" pathway deals with attention, localisation, and tracking of scene objects. We have three modules implementing various functions of the dorsal pathway: shape-based image pre-segmentation, stereo disparity and optical flow. These three modules interact in order to produce a rough layout of the scene and guide the attention of the slower ventral pathway for object recognition.

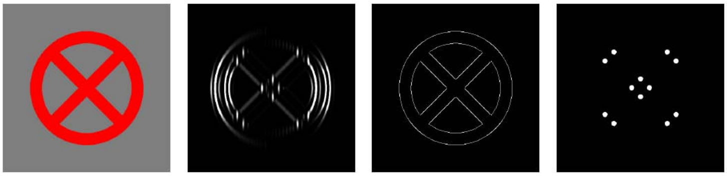

Fig. 3. Results of our V1 model. From left to right: input image, complex cell response (one orientation), detected edges, and detected keypoints, all at one scale.

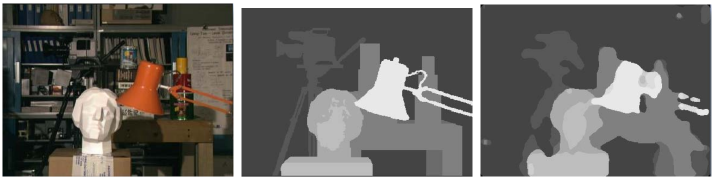

Fig. 4. Disparity Energy Model. Left: an image from the Middlebury dataset. Middle: depth image (ground truth). Right: Depth map produced by our algorithm.

Symmetry and Shape Modelling. We use two methods for detecting salient shapes in images. We model a set of symmetry cells which are active where line and edge segments with compatible orientations are detected at equal distances from the cells. It is known from biology that there are strong cell activations around symmetry axes. We also model a set of salient line and edge segment cells at V2 and V3. Shape grouping cells tuned selective to a set of common geometric shapes are activated when specific segments and symmetry axes are detected at equal distances. Strong activations of populations of these cells correspond to detected shapes. Gestalt-like grouping of salient boundaries is used as an aid for segmentation, attention, and local gist.

Disparity. We use a combination of two main biological disparity estimation algorithms. Some disparity information is available as left-right phase difference of simple cell responses in V1. Early phase-based disparity models gave poor results around discontinuities because the phase of Gabor responses is different for different types of lines and edges [30], but phase can provide exact disparities at line and edge locations as long as the line/edge type is considered during phase calculation. We combine this early phase-based wireframe model with a Disparity Energy Model which works well with large regions. Our DEM consists of about 8000 binocular cells trained using random dot stereograms (see Fig. 4).

Optical Flow. Tracking corner-like features has a long history in computer vision [31]. Our method builds on end-stopped cell responses from V1 [32]. A circular descriptor is calculated for each keypoint by examining simple cell responses around the keypoint and classifying the keypoint as a K, L, T or +

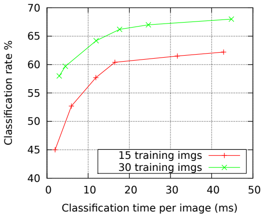

Fig. 5. Recognition performance on Caltech 101 as a function of classification time. Good performance is obtained even at 50 frames per second.

junction. Each keypoint activates the cells of the same class in its neighbourhood so they fire if the same type of keypoint is detected in the following frame. A scale-space tree of keypoints is constructed to group keypoints into objects moving in the same direction, providing rough object segmentation and tracking.

3.3 Recognition

The ventral “what” pathway is in charge of recognition and categorisation. It acts both for extracting the global gist of a scene (scene categorisation) and for accurate recognition of individual objects once a potential object position has been focused on. Complex hierarchical features are extracted but much of the localisation is lost in higher areas like the inferotemporal cortex.

The most influential computational model of the ventral pathway is HMAX [22] in which simple and complex cells are alternated several times, leading to the extraction of stable features. We follow a slightly different approach. We extract keypoints at many scales and extract descriptors around these regions. We then use a Naive Bayes Nearest Neighbour classification algorithm [33] for approximating a MAP classifier. The NBNN classifier only uses a similarity measure, summation and thresholding, which can all be performed by neurons. Our previous work has shown that by using our cortical keypoints instead of a dense descriptor grid, we can significantly cut down on the amount of needed data while maintaining state-of-the-art categorisation performance (see Fig. 5). At the moment, our algorithm uses the SIFT descriptor, which is not completely biologically plausible, but we are currently working on using HMAX-based features constructed from our V1 outputs.

3.4 Short-Term Visual Memory

Our visual system is capable of maintaining a vivid visual description of the scene across saccades and head movements (or pan and tilt action of a stereo camera). This ego-centric representation is somewhere between low-level features and a full semantic scene model (which is maintained in world coordinates). While image representation in V1 is completely retinotopic, there is evidence for a representation which is stable across saccades in higher cortical regions [34].

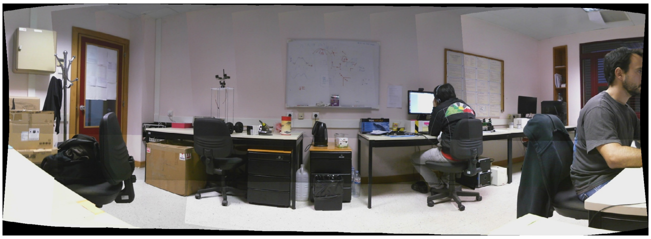

Fig. 6. Panoramic image stitched from 30 camera views. Our V1 keypoints were used together with SIFT descriptors.

Since saccades and movements are inherent in an active vision system, we build a stable representation of V1/V2 responses by stitching together images from different camera views into a panoramic whole (Fig. 6). Since our V1 features are based on wavelets, we can reconstruct the original image with reasonable accuracy from keypoints, lines and edges, so the stitched representation acts as short-term visual memory. Currently, we use standard computational stitching methods with our biological keypoints, but we are working on a fully biological algorithm. We use this intermediate ego-centric memory for robot localisation. Since we are using a pan and tilt unit and not an omni-directional camera (in order to more closely approximate primate vision), locating landmarks requires camera movement and a stable representation is needed to self-localise and estimate a mapping between the current view and the 3D scene model.

3.5 Long-Term Memory and High-Level Reasoning

Our 3D world model is object-based and represented in a 3D coordinate system. For each detected object, we store the position (determined by disparity and triangulation from the short-term visual memory), size, class, shape, primary colours and possibly other features. Each object is updated as the scene changes. In the near future, we will extend this simple model with a biologically motivated dynamical-field represenation [26]. For autonomous robots, we also use a dynamical 2D spatial map of obstacles which is actively updated (using reinforcement learning) and fades with time (see Fig. 9).

High-level reasoning is currently limited to a simple path-finding algorithm in the 3D scene used for our visual SLAM experiments, but we are actively working on sequence learning and task planning. Our 3D model could theoretically be combined with any computational reasoning system such as [11] to infer new information, but reasoning and inference in primates is a field of active research with many unknowns.

3.6 Additional Modules

Non-visual cues can play an important role in active vision, so our system currently also includes two additional modules. The first one is a biological model for binaural sound source localisation, which can detect the direction from which a sound is coming and make a rough guess at the object class. The second module is a dynamical model of the mammalian vestibular system which models head direction cells in the hippocampus in order to estimate the heading direction based on gyroscope readings and can estimate travelled distance based on accelerometer readings.

4 Experiments

Benchmarking generic vision systems is known to be difficult due to a large number of possible scenarios and the work-in-progress nature of most systems. Thus, instead of detailed benchmarks of individual modules, we show our system being used in a number of different scenarios, illustrating the versatility of the system and the ability to leverage different visual processes. The focus is on complex tasks which must be solved by a combination of different modules.

4.1 Real-Time Pose and Gesture Recognition

Figure 7 shows our system detecting and recognising hand gestures in real time. The process begins with the extraction of keypoints, lines and edges (early vision), followed by biologically-inspired optical flow for grouping moving objects together (dorsal stream). Grouped objects are then processed in turn by our object recognition module (ventral stream). The system can successfully distinguish between 5 hand gestures and 5 head gestures at several frames per second.

4.2 Disparity-Based Scene Segmentation

Segmentation of objects in cluttered and textured environments is very difficult, and object detection using sliding windows is expensive. We can apply our system for real-time object recognition in a common robotic table-top scenario (Fig. 8). We start by extracting V1 responses from a complete image (early vision), followed by depth estimation using the Disparity Energy Model (dorsal stream). Disparity produces a rough segmentation corresponding to objects. We then zoom onto each object in turn (foveation), extract keypoints again using our V1 model (early vision), and then perform object recognition using our model (ventral stream). Objects were learned from several views beforehand, but online learning is easy with our approach. The 3D scene model is updated with the size, location, and classes of the detected objects after each frame. Apart from the Disparity Energy Model which is currently being optimised, the system runs in real time.

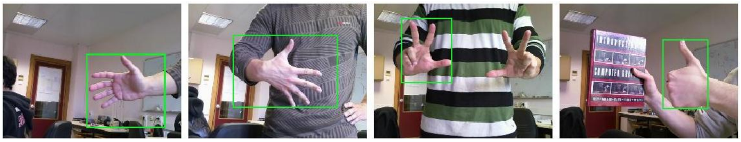

Fig. 7. Optical flow for object detection

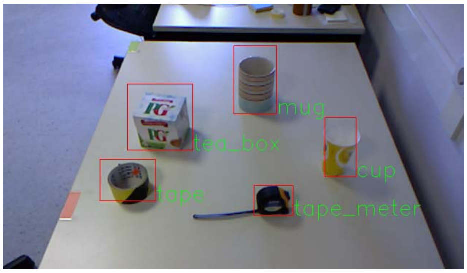

Fig. 8. Real-time object recognition with disparity-based segmentation

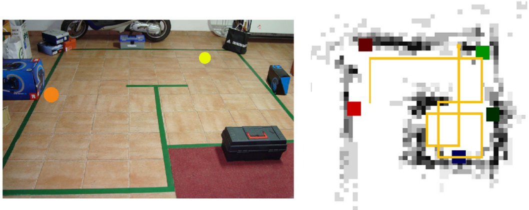

Fig. 9. Robot navigation and SLAM. Left: robot's environment. Right: dynamic world map. Grey areas are detected obstacles which fade with time. Coloured rectangles are detected landmarks overlaid on top of the obstacle map. The yellow line shows the robot's path.

4.3 Vision-Based SLAM

We have updated our robot SLAM algorithm to make full use of the proposed system. The robot can navigate in a simplified environment and look for known objects and use them as landmarks. The process uses coarse-scale V1 keypoints (early vision) to construct a complexity map and guide attention to promising regions (dorsal stream). These regions are then processed sequentially by the object recognition system (ventral stream). The floor is not strongly textured, so areas with many keypoints are assumed to be objects or obstacles. Relative sizes and positions of the recognised landmarks are used to self-localise in world coordinates. A dynamical 2D obstacle map is updated in real time, while landmarks and their coordinates are kept in a separate, object-oriented represenation. Figure 9 illustrates the process.

5 Discussion

We have presented a biologically motivated and plausible system for active vision which combines state-of-the-art biological models into a coherent whole. We have tried to maintain a biological representation as much as possible, with interfaces between modules modelled as maps of neural activations.

Despite years of research on complex computer vision systems, combining different algorithms remains a difficult and unresolved challenge. In the field of biological vision research, this problem is even more pronounced, with a lot of work going into understanding specific areas of the visual cortex, but with little research into combining existing algorithms into a complete vision system. We believe that such work is important, both for understanding the complex interplay between vision subsystems, and for creating practical vision systems.

The system presented in this paper is work in progress. We have evaluated our system in several scenarios and shown that it is capable of solving complex problems which require a combination of visual processes, but there are still many challenges on the road to a complete biological active vision system. Our current work focuses on making all parts of the system biologically plausible and the migration to a larger robot platform which will allow for more challenging experiments in less constrained environments.

Acknowledgement. This work was supported by the EU under the grant ICT-2009.2.1-270247 *NeuralDynamics* and the Portuguese FCT under the grant PEst-OE/EEI/LA0009/2011.

References

1. Tsotsos, J.K.: Analyzing vision at the complexity level. Behav. Brain Sci. 13, 423–445 (1990)
2. Fu, K.S.: Syntactic Pattern Recognition and Applications. Prentice Hall (1982)
3. Leyton, M.: A process-grammar for shape. Artif. Intell. 34, 213–247 (1988)
4. Zhu, S., Mumford, D.: A Stochastic Grammar of Images. Foundations and Trends in Computer Graphics and Vision. Foundations and Trends in Computer Graphics and Vision. Prentice-Hall (2006)
5. Hanson, A., Riseman, E.: Visions: A computer system for interpreting scenes. In: Computer Vision Systems, pp. 303–333 (1978)
6. Guhl, T.P., Shanahan, M.P.: Machine perception using a blackboard architecture. In: International Conference on Computer Vision Systems (2007)
7. Ommer, B., Buhmann, J.: Learning the compositional nature of visual object categories for recognition. IEEE T-PAMI 32, 501–516 (2010)
8. Li, L.J., Socher, R., Fei-Fei, L.: Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In: CVPR (2009)
9. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE T-PAMI 28, 594–611 (2006)
10. Nagel, H.H.: From image sequences towards conceptual descriptions. Image and Vision Computing 6, 59–74 (1988)
11. Neumann, B., Möller, R.: On scene interpretation with description logics. Image and Vision Computing 26, 82–101 (2008)
12. Maillot, N., Thonnat, M.: Ontology based complex object recognition. Image Vision Comput. 26, 102–113 (2008)
13. Fusier, F., Valentin, V., Bremond, F., Thonnat, M., Borg, M., Thirde, D., Ferryman, J.: Video understanding for complex activity recognition. Machine Vision and Applications (MVA) 18, 167–188 (2007)
14. Heitger, F., Rosenthaler, L., von der Heydt, R., Peterhans, E., Kuebler, O.: Simulation of neural contour mechanisms: from simple to end-stopped cells. Vision Res. 32, 963–981 (1992)

15. Hansen, T., Neumann, H.: Neural mechanisms for the robust representation of junctions. Neural Computation 16, 1013–1037 (2004)
16. Tsotsos, J.: Neurobiological Models of Visual Attention, pp. 229–238 (2003)
17. Fidler, S., Leonardis, A.: Towards scalable representations of object categories: Learning a hierarchy of parts. In: CVPR, Minneapolis (2007)
18. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE, 2278–2324 (1998)
19. Fahlman, S.E., Hinton, G.E.: Connectionist architectures for artificial intelligence. IEEE Computer 20, 100–109 (1987)
20. Fukushima, K.: Neocognitron for handwritten digit recognition. Neurocomputing 51, 161–180 (2003)
21. Do Huu, N., Paquier, W., Chatila, R.: Combining structural descriptions and image-based representations for image, object, and scene recognition. In: IJCAI, pp. 1452–1457 (2005)
22. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Object recognition with cortex-like mechanisms. IEEE T-PAMI 29, 411–426 (2007)
23. Milford, M., Wyeth, G.: Mapping a suburb with a single camera using a biologically inspired slam system. IEEE Transactions on Robotics 24, 1038–1053 (October)
24. Siagian, C., Itti, L.: Biologically inspired mobile robot vision localization. IEEE Transactions on Robotics 25, 861–873 (2009)
25. Erlhagen, W., Bicho, E.: The dynamic neural field approach to cognitive robotics. Journal of Neural Engineering 3, 36–54 (2006)
26. Zibner, S.K.U., Faubel, C., Iossifidis, I., Schöner, G.: Dynamic neural fields as building blocks for a cortex-inspired architecture of robotic scene representation. IEEE Transactions on Autonomous Mental Development (in print 2013)
27. Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., Rodriguez-Sanchez, A.J., Wiskott, L.: Deep hierarchies in the primate visual cortex: What can we learn for computer vision? IEEE T-PAMI 99, 1 (2012) (in print)
28. Felleman, D.J., Essen, D.C.V.: Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex, 1–47 (1991)
29. Rodrigues, J., du Buf, J.: Multi-scale keypoints in V1 and beyond: Object segregation, scale selection, saliency maps and face detection. BioSystems 86, 75–90 (2006)
30. Frohlinghaus, T., Buhmann, J.M.: Regularizing phase-based stereo. In: ICPR, vol. 1 (1996)
31. Shi, J., Tomasi, C.: Good features to track. In: CVPR, pp. 593–600 (1994)
32. Farrajota, M., Saleiro, M., Terzic, K., Rodrigues, J., du Buf, J.: Multi-scale cortical keypoints for realtime hand tracking and gesture recognition. In: Proc. 1st Int. Workshop on Cognitive Assistive Systems, Vilamoura, pp. 9–15 (2012)
33. Boiman, O., Shechtman, E., Irani, M.: In Defense of Nearest-Neighbor Based Image Classification. In: CVPR, Anchorage (2008)
34. Turi, M., Burr, D.: Spatiotopic perceptual maps in humans: evidence from motion adaptation. Proc. Biol. Sci. 1740, 3091–3097 (2012)