Proc. of the International Conference on Artificial Neural Networks (ICANN), Madrid, Spain, pp. 795-800 (2002)

Saliency Maps Operating on Stereo Images Detect Landmarks and their Distance

Jörg Conradt, Pascal Simon, Michel Pescatore, Paul F.M.J. Verschure

Institut of Neuroinformatics, ETH / University Zurich
Winterthurerstrasse 190, CH-8057 Zürich
{conradt, psimon, michelp, pfmjv}@ini.phys.ethz.ch
http://www.ini.ethz.ch

Abstract. We present a model that uses binocular visual input to detect landmarks and estimates their distance based on disparity between the two images. Feature detectors provide input to saliency maps that find landmarks as combinations of features. Interactions between feature detectors for the left and right images and between the saliency maps enables corresponding landmarks to be found. We test the model in the real world and show that it reliably detects landmarks and estimates their distances.

1 Introduction

Animals need a spatial representation of their environment to navigate and return home. There exist neurons in the hippocampus that are only active in a small part of the environment, the so-called place field [1]. These neurons provide a spatial representation, which is built on a variety of sensory information. However, the dominant stimuli are distant visual landmarks [2]. This paper proposes a model that detects landmarks in natural scenes using a stereo video system and estimates their directions and distances.

A number of biologically inspired algorithms exist that find salient objects in monocular camera images, some of which use saliency maps (SMs) [3]. While it is possible to extract a small number of distance cues from monocular images, e.g. landmark size, occlusions, etc., these approaches do not estimate distance.

In contrast, multiple images taken from different positions contain an important distance cue, the differences in the landmark's position in each image, called disparity. It turns out that detecting the same landmark in both images is a hard problem [4] due to e.g. only partially overlapping visual areas of both cameras, varying lighting conditions, noise in the signal.

The approach we explore in this paper is based on biologically inspired Saliency Maps (SMs) [5, 6] that receive preprocessed input from Feature Detectors (FDs). Interactions between both cameras' FDs and SMs support the detection of corresponding landmarks in both images and allow the estimation of their direction and distance.

2 The Hardware Setup

We use a mobile Koala Robot (K-Team, Switzerland) with a stereo pan-tilt camera system (fig. 1) to perform real world experiments. A standard Personal Computer (Atmel processor, 1400 MHz, RedHat Linux 7.2) controls the robot and the cameras. Both the robot's cameras (type VPC-795) are mounted on pan-tilt devices, allowing each camera to turn in 2 degrees-of-freedom with minimal translation of the camera's optical center. Both cameras can move independently, but their tilt angles are coupled in software. In the PC, two standard PCI framegrabber boards (Hauppage WinTV) acquire the images. To increase the computational efficiency, we reduce the image resolution (768x576) by a factor of 4 in both directions. The landmark detection, and the control of the pan-tilt system and the robot run in real-time on the PC. The system is implemented in MatLab, Version 6 (The Math Works, USA).

3 Description of the Model

3. 1 Feature Detectors

A saliency map based approach to finding landmarks in images consists of multiple processing stages. Firstly, we extract four different channels from every image: red, blue, green and brightness information, with normalized color channels. Every channel is then convolved with a set of feature detectors (FDs) that respond best to particular patterns in the image. The FDs used are circles of different radius, and horizontal, vertical, and diagonal edges (shown in fig. 3). Convolving the FDs with all channels results in 24 independent feature selective maps. To preserve the FDs' polarity, the 24 maps are split into 48 feature maps (FMs), each containing either a map's positive stimulus or the absolute values of a map's negative stimulus.

3. 2 Saliency Map Based Detection of Landmarks

All FMs are used as input to a Saliency Map (SM) that is composed of a competitive neural network as proposed in Amari and Arbib's neural fields approach for modeling cortical information processing [5, 6]. The goal of the network is to take the FDs' spatially localized stimuli as input, have them compete and finally output a winning target. The neurons' activation dynamics in the SM is expressed as

$$\tau \dot{u}(\mathbf{x}) = -u(\mathbf{x}) + S(\mathbf{x}) + h + \sum_{\mathbf{x}'} w_k(\mathbf{x}, \mathbf{x}') \cdot \sigma(u(\mathbf{x}')) \,, \tag{1}$$

with h being the baseline activation level and $\sigma(u)$ controlling the local threshold of activation. $S(\mathbf{x})$ is the cumulative weighted input from the FDs, defined as

$$S(\mathbf{x}) = \sum_n w_n \cdot stimulus_n(\mathbf{x}) \,. \tag{2}$$

Depending on the choice of the parameter *h* and the form of σ and w_k, the activation dynamics of (1) can have various stable equilibrium points. We are interested in a solution which has uniform low activation in regions without stimuli, and which forms a peak of activity at the location of the most significant combination of features. This is achieved by using a smoothed step-function

$$\sigma(u) = 1/(e^{(-cu)} + 1) \tag{3}$$

as the transfer function and an interaction kernel with short-range excitation and a long-range inhibition term H_0

$$w_k(\mathbf{x}, \mathbf{x}') = k \cdot e^{-(\mathbf{x}-\mathbf{x}')^2/\sigma_w^2} - H_0 \,. \tag{4}$$

The values of the constants H_0, *k*, *c*, and σ_w^2 have to be determined based on the magnitude of the stimulus $S(\mathbf{x})$, as outlined in [5]. The SM has found a target if the local activity in the network exceeds a global threshold.

3. 3 Finding Corresponding Landmarks in Stereo Images

We will use the offset between an object’s positions in both images to estimate its distance from the cameras. To find landmarks in two images, we use independent FMs and SMs for visual input from a left and a right camera. Ideally, the first targets found independently in both images would correspond to each other, as would the second, the third, and so on. However, false pairing will occur due to e.g. the shifted visual fields of both cameras, noise in the signal, or varying lighting conditions.

Interactions between Left and Right Processing Streams

The detection of corresponding targets despite a potentially substantial offset is easier at the level of the FMs, where the landmarks’ features are separate. To enhance potentially corresponding stimuli, we introduce a coupling between FDs on the left and right processing streams that respond to the same feature (e.g. small bright circles). Activity in one FM excites a one-sided Gaussian shaped region in the corresponding FM of the other camera, whereas the absence of activity inhibits a small region (fig. 2). Corresponding landmarks lie at similar vertical position in both images, such that the vertical extension of the interaction kernel is minimized to avoid false pairing. In contrast, the horizontal width of the interaction adapts to a landmark’s expected offset in both images, which is reduced during consecutive iterations of the algorithm as described below.

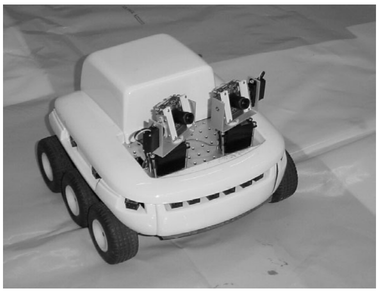

Fig. 1. The mobile Koala robot with stereo pan-tilt cameras.

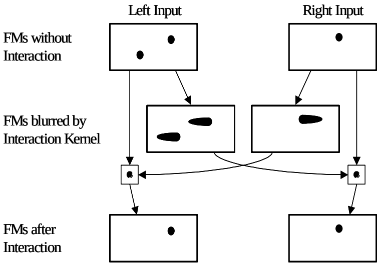

Fig. 2. Interactions between corresponding FMs enhance corresponding stimuli and suppress stimuli without matches.

Additionally, both SMs interact to enhance activity that originates from corresponding stimuli but different FMs. Activity in one SM excites the other within a one-sided Gaussian-shaped region, centered on the same position.

Determining Landmarks

If neurons in both saliency maps at similar vertical locations exceed a threshold, the positions in both images define the new target. The required adjustments in camera pan and tilt angles to center the target are estimated by linear approximation. The peaks of activity in both SMs are then shifted to the center of the map to adjust for the cameras' rotation.

If a target was found outside of the image center, the SMs continue processing input to increase the accuracy in estimating the target's position. In such a case, we introduce a short-term attentional bias by increasing the connection weights w_n between the SMs and those FMs that contributed to the target. Additionally, the width of the interaction Gaussian is reduced, as the positions of the target will be nearer in both images.

If the target is already close to the images' center when detected, a new landmark is declared found. The landmarks' coordinates and the cameras' pan and tilt angles are stored. Finally, a strong negative activity is introduced in both saliency maps at the position of the landmark, generating an inhibition of return [5], allowing other targets to be detected in subsequent iterations. Additionally, the weights w_n between the FMs and the SMs are reset to initial values.

3. 4 Estimating a Landmark's Direction and Distance

When the SMs have found a landmark, its direction and distance are estimated. A neural net maps the 7 given coordinates (the landmark's x and y position in both images, 2 camera pan angles, and one common camera tilt angle) to the landmark's direction and distance.

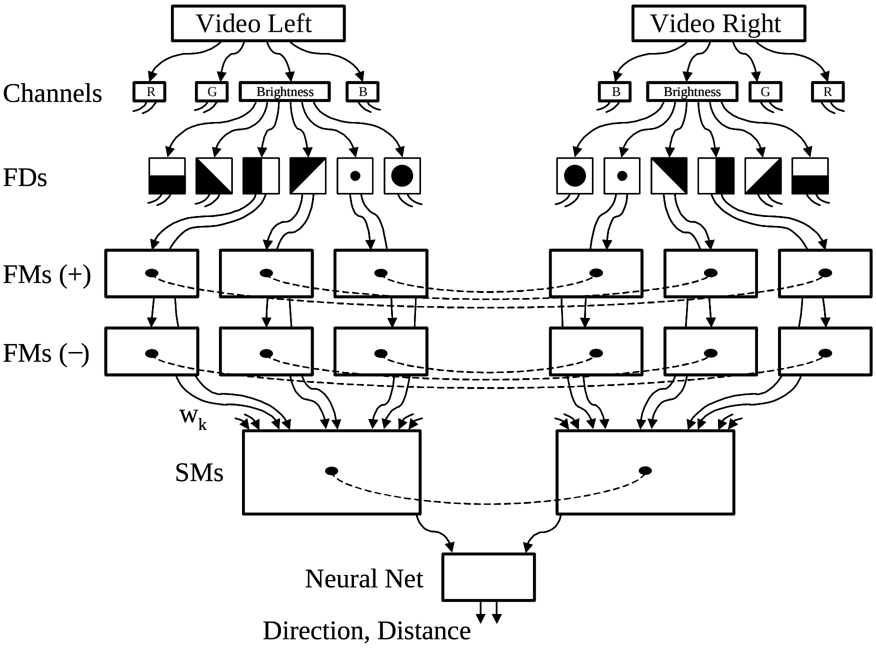

Fig. 3. From top to bottom: Each video image is split into four channels, each convolved with 6 FDs. The resulting 48 FMs on each side (only six shown here) serve as weighted stimulus for a SM that detects landmarks. The SMs provide input to a neural net estimating direction and distance to the landmark. Interactions between corresponding FMs belonging to the left and right stream and between the SMs are indicated by dotted lines.

4 Results and Discussion

The number of landmarks detected in a stereo image varies dramatically with the image content. The algorithm typically detects 16-20 corresponding targets in our office environment, using the whole pan-tilt range of 150 horizontal and 45 vertical degrees. These targets do not all represent distinct landmarks, as some targets fall on the same object. A result of the detection algorithm for a single pair of stereo images is shown in figure 4. In this scene, nine distinct targets are found. High contrasts targets are detected first, here a green key-ring mascot on the shelf, followed by a blue light bulb in the lower right-hand corner. Successively, the monitor, the computer rack, the oscilloscope, the folder, and the wooden lamp-holder are detected. Finally, the cables at the very top are found. Then, the algorithm returns to the previously detected objects, but attends to those found first more frequently. The error in estimating a landmark's distance is below 10% within six meters.

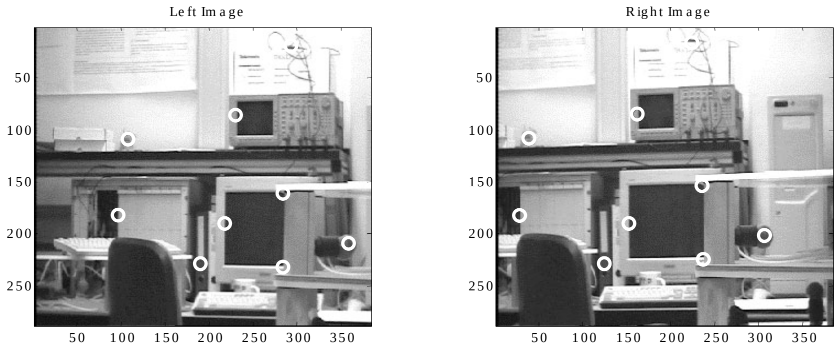

Fig. 4. Examples of matching target pairs denoted by white circles are shown. Here, the pan-tilt system was disabled and the algorithm searched for landmarks in static images.

Random dot stereograms, for example, show that humans recognize depth based on disparity without recognizing objects [4]. This inspired us to implement interactions between FMs, rather than between SMs only, which turned out to it increases the reliability in matching points.

A number of algorithms for landmark detection and matching have been proposed, e.g. [7-9], but they are either computationally expensive or biologically implausible. The algorithm presented here detects landmarks and their distance in natural environments in a computationally inexpensive and biologically inspired way. The model runs in real time on a mobile robot. We will use it to provide the directions and distances of landmarks for a model of hippocampal place fields, for use in robot navigation.

References

[1] J. O'Keefe and L. Nadel, *The Hippocampus As a Cognitive Map*: Oxford University Press, 1978.

[2] P. J. Best, A. M. White, and A. Minai, "Spatial Processing in the Brain: The Activity of Hippocampal Place Cells," *Annual Reviews of Neuroscience*, vol. 24, pp. 459-86, 2001.

[3] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shift of visual attention," *Vision Research*, vol. 40, pp. 1489-1506, 2000.

[4] I. P. Howard and B. J. Rogers, *Binocular Vision and Stereopsis*, vol. 29. New York, Oxford: Oxford University Press, Clarendon Press, 1995.

[5] S. Amari, "Dynamics of pattern formation in lateral-inhibition type neural fields.," *Biological Cybernetics*, vol. 27, pp. 77-87, 1977.

[6] S. Amari and M. A. Arbib, "Competition and cooperation in neural nets," in *Systems Neuroscience*, Metzler, Ed. New York: Academic Press, 1977, pp. 119-165.

[7] P. E. Trahanias, S. Valissaris, and T. Garavelos, "Visual Landmark Extraction and Recognition for Autonomous Robot Navigation," presented at International Conference on Intelligent Robots and Systems, Grenoble, France, 1997.

[8] S. Livatino and B. C. Madsen, "Autonomous Robot Navigation with Automatic Learning of Visual Landmarks," presented at International Symposium on Intelligent Robotic Systems, Coimbra, Portugal, 1999.

[9] G. Bianco, A. Zelinsky, and M. Lehrer, "Visual Landmark Learning," presented at International Robots and Systems, Takamatsu, Japan, 2000.