Developmental Stereo Vision

Arden Knoll and Jacob Honer

MICHIGAN STATE UNIVERSITY GENISAMA

Autonomous robots rely on many sensors, such as lidar and cameras, to become spatially aware and navigate their environments. While Lidar has become a popular option, it is expensive, large, and lacks the rich information that is available to cameras. Stereo cameras could be low-cost, highly mobile, solutions if computer vision algorithms could adequately utilize the wealth of information provided by the cameras. Traditional stereo vision algorithms, however, are brittle, as they struggle to handle complex distortions provided by real-world image pairs such as lighting differences between cameras, differences between viewing angles, and occlusion. Neural networks can address these problems by learning real-world representations and have found success in many visual tasks (including disparity detection). However, neural networks are used similarly to traditional stereo-disparity detection algorithms, where they generate a disparity value for each pixel and provide a depth map as input for another algorithm. Developmental Networks (DNs) are capable of attention, meaning they are not only able to predict disparity, but can also determine where to look. Furthermore, the DN can act as a full vision system by learning tasks that require both monocular and binocular visual information. This work provides a method for learning stereo-disparity detection and for integrating binocular cues with monocular cues. We discuss the inner workings of the DN algorithm in terms of disparity detection and explore how the DN can learn to implicitly detect disparity through unsupervised updates.

INTRODUCTION

DNs are capable of attention and can incorporate both binocular and monocular information through their learning processes. Therefore, DNs lend themselves to be useful tools in autonomous navigation tasks. In this work, we explore the DNs ability to predict stereo-disparity and integrate binocular and monocular cues. Stereo-disparity is a shift between corresponding pixels in a Left and Right binocular image. Here, we do not require the DN to generate its own depth map, as that process would require too many computations to approach real-time performance. Instead, we propose a method where, through attention, the DN predicts a singular disparity value of the closest object in frame and can decide further navigation actions based on the monocular pattern of the image. Past work has demonstrated stereo-disparity detection via a DN on simulated image shifts [1].

Attention in the DN

The DN acts as a statical probability model that selects disparity using the generalized supervised weights (see Fig 3) from the DN's learning epochs [2]. We also propose two methods internal to the DN that help with the DNs statistical computations: volume dimension and subwindow voting [3]. Volume dimension adds a new dimension to each neurons input. This input is set to high when the other normalized inputs are low. This helps the DN attend to strong textures, reducing the attention given to weaker textures in computation. Subwindow voting leads to an output disparity that is weighted according to each subwindow's analyzed certainty. This means that weak textures have less influence in "voting" for the correct disparity compared to stronger textures, hence, the DN "pays attention" to the stronger textured receptive fields. Attention is one of the major novelties of using a DN for stereo-disparity detection.

Integration of Monocular and Binocular Cues

Neuron's overtime become tuned to a specific binocular pattern, or specific shift in their two input images. This pattern can be used as neurons "vote" to produce a disparity. Concurrently, the monocular patterns can be used to inform other actions. Therefore, one neuron represents both binocular and monocular context, allowing the DN to seamlessly integrate information from both [4].

Implicit Learning

Is the idea that the DN learns continuously (frame by frame), constantly updating its bottom-up weights, whether supervised or not. Therefore, an individual neuron within the DN will update on similar features so that its bottom-up weights become more generalized, and the neuron becomes tuned to a specific disparity and monocular pattern. Thus, the DN can continuously learn and generalize through passive updates (practice mode) [4].

Performance Equivalent DNs

Another important aspect of a DN is that each DN is performance equivalent. The amnesic learning nature of a DN means that on each neurons first firing, the learning rate is one and the retention rate is zero, hence, the randomly initialized weights are immediately forgotten. This avoids any need to select the best performing DN of many randomly initialized DNs [2].

EXPERIMENTS

The main design considerations for stereo-disparity detection using the VCML-100 include the following mutually conflicting constraints: (1) real-time speed (2) cost, and (3) mobility. Accounting for the above constraints, we chose the following DN parameters. (1) The input is grabbed from a small, 135×135 pixel, mask on both the left and right images (red square in the figure below). (2) The hidden neurons are grouped into nine columns (inhibition zones), each column sharing the same initial receptive fields. (3) The input image is divided into 3×3 non-overlapping subwindows, of size 45 × 45 pixels, where each subwindow is the receptive field for the neurons in its respective column. (4) The number of hidden neurons is limited. The figure below shows this DN architecture.

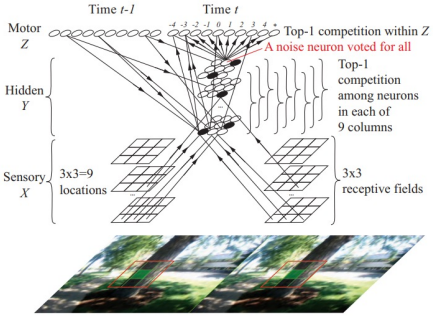

Fig 1. Visualization of our DN architecture.

Such a choice of parameters affects performance, but a practical system must be real-time with limited computational power (i.e., a smart phone). To help achieve real-time speeds, we designed a more optimal version using the GPU and multiple CPU cores of the Android phone. The improved rates of the more optimal version can be seen in table 1.

This work consists of two experiments. The first is meant to demonstrate how the DN can learn to detect the disparity of the nearest object through direct supervision. The second extends this idea by training on a larger amount of data and letting the DN learn disparity detection implicitly through passive updates (practice without supervision).

For the first experiment, we used a sequence of 1400 stereo images, recorded and labelled in real-time from the natural world. The DN was trained on frames from odd indices and tested on even frames. For this experiment, every DN had 100 neurons per column (900 total), top-8 competition in each of the 3×3 = 9 columns, and a global top-3 competition in the motor area.

The data in the second experiment consisted of three stereo image sequences (Nav-1, Nav-2, Nav-3) collected in real-time from an outdoor walkway setting like the one in Fig 1. Nav-1, Nav-2, and Nav-3 contained 6502, 6686 and 5508 frames, respectively. We let the DN live through 14 sessions, as shown in Table II. The DN has a growth rate with which half of the 750 neurons per column (750 × 9 = 6750 total hidden neurons) are activated in session 1 and half session 3.

RESULTS

Experiment 1

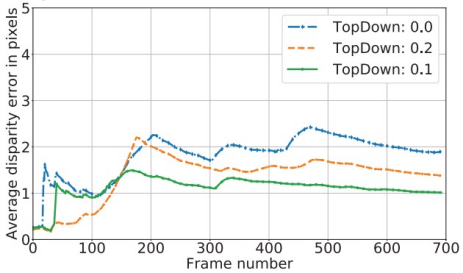

Fig 2. Experiment 1 performance data

The incrementally computed average errors from time zero of all disjoint tests are shown in Fig 2. When the top-down weight was 0.1, the average error was the smallest, reaching 1.0 pixels.

Version	Training rate	Frozen-testing rate
CPU	0.96 Hz	1.00 Hz
GPU	7.16 Hz	10.47 Hz

Table 1. Experiment 1 performance data

Table 1 shows the increased update rate achieved once the heavy computations of the DN were performed using the GPU and multiple CPU cores of the Android phone. The GPU version approaches the real-time speeds necessary for a real-world system.

Experiment 2

Session	Nav	Mode	Disparity	Heading	Stop/Go
1	1-even	Motor-S	0.00 px	0%	0%
2	1-odd	Frozen	1.27 px	10%	2%
3	2-even	Motor-S	0.00 px	0%	0%
4	2-odd	Frozen	1.43 px	15%	3%
5	1-odd	Frozen	2.19 px	17%	2%
6	3-odd	Frozen	2.74 px	20%	3%
7	3-even	Practice	2.71 px	21%	3%
8	3-odd	Frozen	2.70 px	23%	4%
9	3-even	Practice	2.67 px	24%	3%
10	3-odd	Frozen	2.70 px	23%	4%
11	3-even	Practice	2.68 px	26%	3%
12	2-odd	Frozen	1.65 px	15%	2%
13	1-odd	Frozen	2.31 px	21%	3%
14	3-odd	Frozen	2.67 px	26%	3%

Table 2. Experiment 2 performance data

In table 2, session is the epoch number, Nav is what sequence was used, Mode is the DN mode used (motor supervision, frozen, or practice), Disparity is the average disparity error over the whole session in pixels, heading is the classification error of what direction the user should head (left, straight, right), and stop/go is the classification error of whether to stop or continue (go).

Table 2 shows that the DN self-supervising in practice mode improved performance. This can be explained through practice mode allowing the weights of the DN to further update and generalize over epochs, allowing for a smoother transitions and an increased certainty in the statical computations of the DN.

The DN not only performed well on disparity, but also predicted the direction the user should be heading in and whether the user should stop/go with low error.

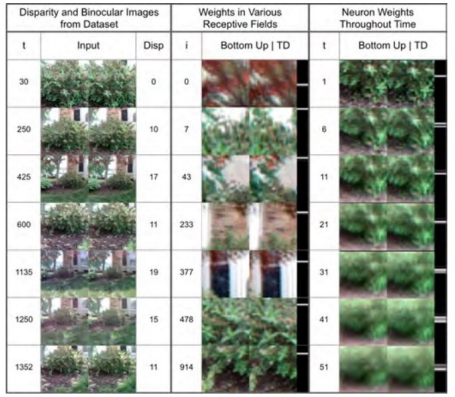

Fig 3. Weight visualization data from experiment 1

CONCLUSIONS

This work demonstrated the DNs capabilities to not only detect disparity, but also to integrate binocular and monocular patterns to accomplish a more complex task. We also demonstrated the benefits of implicit learning on feature generalization.

The DNs are potentially monumental tools for perception systems, as they can give autonomous robots a sense of spatial awareness. DNs are robust and general purpose, lending themselves to a wide variety of applications. Future work might expand on this work to include more modalities and more complex visual tasks.

REFERENCES

[1] M. Solgi and J. Weng. Developmental stereo: Emergence of disparity preference in models of visual cortex. IEEE Trans. Autonomous Mental Development, 1(4):238–252, 2009.

[2] J. Weng. Natural and Artificial Intelligence: Introduction to Computational Brain-Mind. BMI Press, Okemos, Michigan, second edition, 2019.

[3] J. A. Knoll, V. N. Hoang, J. Honer, S. Church, T. H. Tran, and J. Weng. Fast developmental stereo-disparity detectors. In Proc. IEEE International Conference on Development and Learning and Epigenetic Robotics, pages 1–6, Valparaiso, Chile, Oct. 26-27, 2020.

[4] J. A. Knoll, J. Honer, S. Church, and J. Weng. Optimal Developmental Learning for Multisensory and Multi-Teaching Modalities. In Proc. IEEE International Conference on Development and Learning, to be presented August, 2021.