A visual attention mechanism for autonomous robots controlled by sensorimotor contingencies

Alexander Maye*, Dari Trendafilov†, Daniel Polani† and Andreas K. Engel*

* University Medical Center Hamburg-Eppendorf
Dept. of Neurophysiology and Pathophysiology
Martinistr. 52, D-20246 Hamburg, Germany
{a.maye,ak.engel}@uke.de

†University of Hertfordshire
School of Computer Science
Hatfield, UK
{d.trendafilov,d.polani}@herts.ac.uk

***Abstract*— Robot control architectures that are based on learning the dependencies between robot's actions and the resulting change in sensory input face the fundamental problem that for high-dimensional action and/or sensor spaces, the number of these sensorimotor dependencies can become huge. In this article we present a scenario of a robot that learns to avoid collisions with stationary objects from image-based motion flow and a collision detector. Following an information-theoretic approach, we demonstrate that the robot can infer image regions that facilitate the prediction of imminent collisions. This allows restricting the computation to the domain in the input space that is relevant for the given task, which enables learning sensorimotor contingencies in robots with high-dimensional sensor spaces.**

I. INTRODUCTION

Sensorimotor contingency theory [1] captures the idea that the regularities between actions and ensuing sensory signals define perceptual experience rather than the properties of these signals per se. We argue that this action-based approach to cognition holds great potential for making progress in the development of artificial cognitive systems, however only few robot control architectures to day implement this concept. One problem is the size of the memory needed to learn patterns in the sensorimotor flow when sensor spaces are large, which is the case, for example, in robots that are equipped with cameras or artificial skins. This problem can be tackled by limiting the computations to the subset of sensory elements that is relevant for the given task and context. In essence this constitutes an attentional mechanism that fulfills a similar function in biological agents.

In this paper we suggest that the robot can learn autonomously to which regions in the visual field of its camera it should pay attention in order to avoid collisions. The main idea is to learn the relevance of the visual sensorimotor contingencies (SMCs) for predicting collisions in the near future. We analyze the optic flow that the movements of the robot induces as a paradigmatic example of SMCs which are specific for the visual modality. When the robot moves in different directions, the motion flow field shows patterns which are characteristic for the movement direction and the spatial layout of the scene.

The approach is inspired by a previous study in which we showed that the correlation between each sensory channel and a utility function, which reflects the suitability of the current state for solving the task, can be used to constrain the evaluation of the sensory input to the channels that are the most relevant in the given context [2]. This leads to, among other advantages, faster learning of successful actions and differentiation between sensory modalities. Here we extend this approach by an information-theoretic analysis of the relation between the motion flow and subsequent collisions. In particular we suggest that regions in the visual field with high mutual information (MI) between the motion flow and the collision state in the subsequent time step are relevant for avoiding collisions.

Analysis of motion flow is a popular approach in computer vision for functions like motion detection, scene segmentation, three-dimensional space reconstruction, estimation of time-to-collision and stereo disparity measurement [3]. Several methods for predicting imminent collisions from motion flow exist, e.g. by flow field divergence [4], [5], computation of egomotion parameters [6], models of the motion flow [7], or population-coded image velocity features [8]. In general, these approaches analyze the images in a sequence completely to determine the location where collisions need to be expected. Our idea is that there are statistical regularities in the relation between these locations and the direction in which the agent is moving, and we attempt to infer these regularities by analyzing the MI between the motion flow and subsequent collisions.

A variety of approaches to calculate motion flow has been developed [3]. In contrast to the way how motion flow is employed in computer vision applications, we do not attempt to estimate the veridical flow here. Similar to the elementary motion detectors in the insect brain [9], the simple pixel-correlation-based approach we use responds to the spatial information in the scene as well as to textural information of surfaces. The dependency of the flow field from the patterning is typically considered as noise which has to be minimized. Computational models of the motion flow analysis in insect brains suggest though that animals may make use of this patterning information [10]. We selected this approach also because it allows, at least in principle, the estimation of motion flow in real-time and by a neuronally

This work was supported by the H2020-641321 socSMCs FET Proactive project `socSMCs.eu`

plausible architecture.

Information theory provides powerful tools for analyzing the coupling between an agent and its environment. The study by [11], for example, investigated the information flow between sensors and actuators in a quadruped robot which was moving on different ground materials. The information flow from the motors that control the robot's legs to the sensors (measuring, e.g., joint angles, forces, accelerations) was analyzed using transfer entropy (TE). They found that the informational coupling was shaped by the physical interaction between the body and the different ground materials. A quantitative analysis of the transfer entropies revealed that different sensory channels have different predictive capabilities for upcoming sensory states; therefore, they differ in their utility to the agent. Channels with high predictive capabilities should receive more attention than channels with low predictability.

The standard model for visual attention determines regions of interest from features in the images [12], [13]. In this model, various features are computed across a given image, and regions with high feature values are selected for subsequent visual search. In contrast to this sensor-based, bottom-up mechanism, our model is inherently action-based and combines bottom-up with top-down elements. This is achieved by associating relevance maps (computed by mutual information) with SMCs, which are characteristic for an interval of sensorimotor interaction (see [14] for a consideration of timescales). While the selection of a relevance maps depends on the action context, which can be considered as a top-down attention mechanism, the shapes of the different maps is a result of the bottom-up information that is captured by the SMCs.

II. Methods

All data were acquired during our previous studies on the differentiation between sensory modalities [2] and prediction and planning using SMCs [15]. In total, this data set comprises approx. 150.000 epochs of 500ms duration each, corresponding to a runtime of about 20hrs. We had not analyzed the video stream data so far, which is the focus of the current study. All details about the experimental setup and the methods are given in the cited studies. Here we focus on the description of details that are relevant for the investigated topic.

A. Experimental setup

The setup consisted of a Robotino robot (Festo Didactic, Esslingen, Germany) which was freely roaming a rectangular environment of about 1x2m in the laboratory (see Fig. 1). Among other sensors that are not used in this study, the robot was equipped with a collision detector and a rigidly mounted webcam. The collision detector was implemented by an air tube around the circular periphery of the robot and therefore could not give information about the location of the contact. The webcam streamed images at a frame rate of about 5Hz to a computer, where image sequences were processed. The environment consisted of the irregularly patterned linoleum floor, a white locker and a white door at the short sides of the arena, and patterned wallpaper and a cardboard wall at the long sides. The environment was illuminated by regular neon lights at the ceiling. No attempt was made to optimize the environmental conditions with respect to the data analysis.

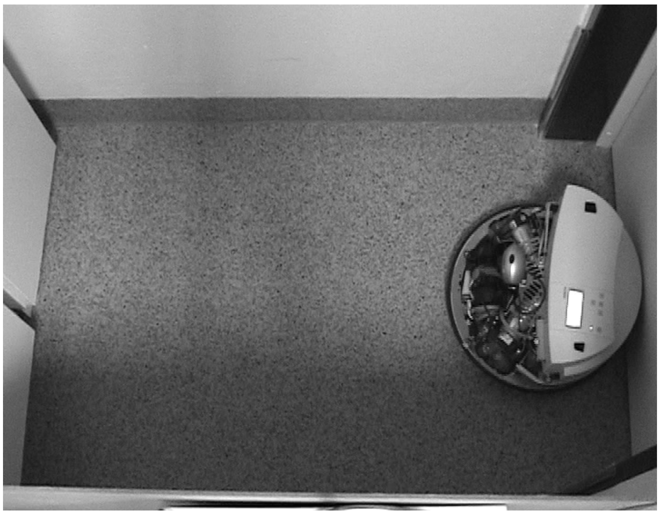

Fig. 1. View on the robot and its environment.

Although the Robotino features an omnidirectional wheel drive, movements were constrained to the 4 cardinal directions: forward, backward, left and right. The robot could change direction after each epoch, that is, every 500ms. The behavior of the robot was controlled by an action selection schema for maximizing a utility function as described in [15]. The utility function is a superposition of the robot's acceleration, current energy consumption and collision state, and attained its maximum for straight movements between the walls without collisions. Apart from the utility function, no pre-existing knowledge, instructions or automatic behaviors were given to the robot. The only task was to maximize the utility function. To this end, the robot had to gather physical experiences with collisions and explore appropriate actions to escape them. The data therefore allow searching for regularities in the sensorimotor flow which are predictive of unfavorable events in terms of utility. As the collision state had the strongest influence on the utility function, and the motion flow is not expected to predict upcoming accelerations and current consumptions, we use only the collision state in the analyses here.

B. Data processing and motion flow estimation

Data from all sensory channels were sampled at 5Hz and processed in real-time. The only exception was the video stream, which was preprocessed in real-time, but analyzed offline. The raw camera images were resampled to a resolution of 40×30 RGB pixels.

Motion flow was estimeated from two consecutive images by determining the offsets in x and y direction by which an image block of 5×5 pixels had to be shifted to give maximum correlation with the corresponding block in the previous image. The estimated flows from each image pair were averaged across the whole epoch, resulting in one estimate of the motion flow per epoch.

The value of the collision time series was 1 if there was a collision during an epoch and 0 otherwise.

C. *Information-theoretic analysis*

Our primary motivation for an information-theoretic analysis of the relation between visual data and utility values came from the difficulty to apply a Pearson correlation coefficient, like we did in our previous study [2], to the vectors in the motion flow field. Here we applied standard mutual information in order to measure correlation. Unlike the study described in [11], which employed TE to estimate information flows between actuators and sensors, we use MI in our study for two reasons: The first reason is simplicity. Without simplifying assumptions, robust estimation of information-theoretic functionals usually requires a large number of data samples. Accurate estimation of entropy-based measures like TE is notoriously difficult. Every method has its own free parameters, and there is no consensus on an optimal way of estimating TE from a data set [16]. The second reason is that measuring information-transfer-type quantities requires prior establishing the presence of and quantifying causal relationships, as argued in [17], by applying, for example, information flow techniques, as suggested in [18]. Revealing causal effect necessitates some type of perturbation or intervention of the source, so as to detect the effect of intervention on the destination. Attempting to infer causality without doing so, leaves one measuring correlations of observations, regardless of how directional they may be. All methods that rely on observing time series only have difficulties separating real causal effects from spurious correlations. TE, alike, can be viewed as conditional MI; therefore, it is still a measure of observed correlation rather than of direct effect. Having established the causal relationships, we can subsequently analyze information transfer. However, in order to be genuinely interpreted, TE should only be measured for causal information sources that contribute to the given destination. Beyond these sources, it only measures correlations that don't directly contribute or transfer information into the computation that determines the next state of the destination. Establishing the causal relationship between motion flow and collision states is left for future work.

In order to compute MI we defined a discrete stochastic model representing the random variables in our data set. We used a standard binning approach to estimate the joint probability distributions, from which we derived the value of MI. We discretized the real-valued average motion flow by employing the following simple method without loss of precision. The flow vectors, computed from two consecutive frames, are integer-valued in the range $[-3, 3]$. Since the average flow for each epoch is only over four elements, the number of different values it can assume is finite. By adding the constant offset (3,3) to and multiplying all vectors by 100, we converted them to a non-negative integer-valued estimate of the motion flow for each epoch. Furthermore we combined the two elements of each vector into a single value, by 'shifting' the x-component three magnitudes to the 'left' (by multiplying it with 1000) and adding the y-component. That is, the integer representation of the motion flow $f(t)$ in epoch t was computed from the vectors $(x(t), y(t))$ as $f(t) = 1000\,\text{fix}(100(x(t)+3)) + \text{fix}(100(y(t)+3))$.

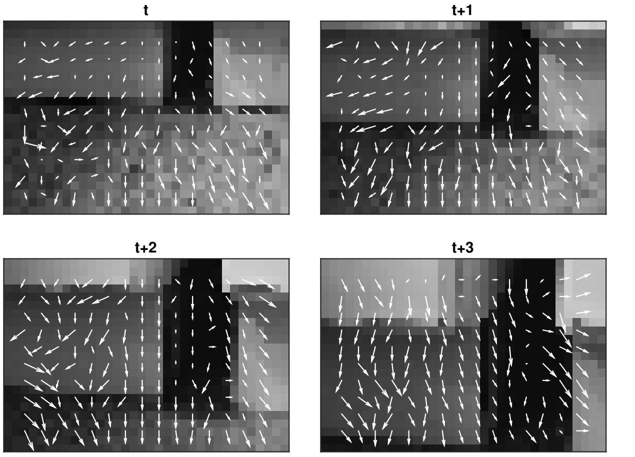

Fig. 2. Example image sequence for a forward movement towards the wall closet on the left side of Fig. 1. White arrows show the local motion flow that is computed from two consecutive frames.

III. Results

We analyzed episodes in which the robot moved in the same direction for 5 consecutive epochs. This corresponds to a traveling time of 2.5s and a distance of about 40cm. The data set contained around 3.000 of such episodes for each of the 4 directions.

The motion flow depends on the robot's movement as well as on the spatial configuration of the scene, the depth of object surfaces in particular. Obviously the magnitude of flow vectors is larger for close than for distant walls, which allows predicting collisions from information in the flow field. This can be seen in the example in Fig. 2, where the motion flow on the ground just in front of the robot is larger than at the surface to which the robot is moving. The example also visualizes the noise in the estimated motion flow that is generated by the simple correlation-based method we employ. Regions with dim illumination seem to be affected in particular. As the robot could change the movement direction only once per epoch, we averaged the motion flow obtained from the individual frame pairs during an epoch in order to arrive at a single estimate for this epoch in which this noise is reduced. The estimates for each epoch were then analyzed together with the collision state of the corresponding subsequent epoch.

In Fig. 3 we present the average motion flow over all epochs for each of the 4 movement directions. They exhibit in general the expected patterns, with distortions that result from the idiosyncrasies of the camera optics, the environment, as well as the movements. These flow fields visualize the systematic dependency of the shift in the visual input from the executed movement, which is what we consider a visual SMC.

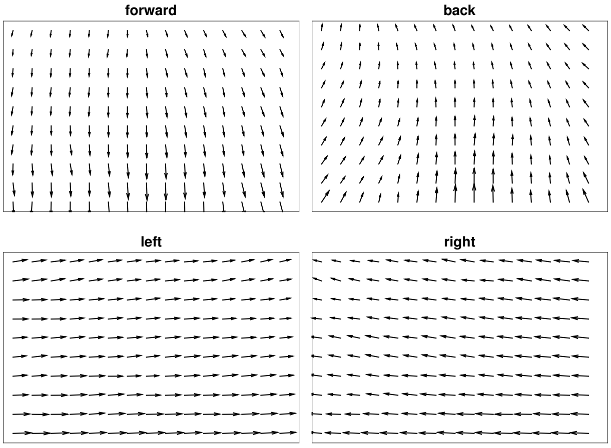

Fig. 3. Grand average motion flow for the robot's four movement directions.

We analyzed the MI between the motion flow in individual epochs and the subsequent state of the collision detector. Figure 4 shows the MI for each point in the visual field and for the four movement directions. For forward movements, regions at the bottom center and at the left and right lower half of the visual field feature high MI values. During backward movements, only the region at the bottom center is the most informative, and absolute MI values are lower compared to forward movements. For left- and rightward movements, MI maps are less structured and show the highest values at the bottom of the visual field. The center of gravity is right of vertical axis for leftward movements and vice versa for rightward movements. In terms of absolute values of MI, it has to be noted that for movements to the right, they are higher than for all other directions. The ranges for forward and leftward movements are similar, and backward movements feature the smallest values.

IV. Discussion

The regularities of the motion flow for different movements are mainly driven by the laws of perspective distortions. The imperfect symmetries and the jitters in the flow fields indicate however that these regularities are modulated by the properties of the embodiment of the individual robot that we used for this study. An interesting finding is that the motion flow can help with forecasting collisions even when the robot moves backward, although to a lesser extent than when moving in the other directions. For explaining this property one may consider the situation of swimming backstroke, where the visual information allows predicting when to turn although the respective wall of the pool is in the opposite direction of the viewing field. Clearly this prediction capability requires some experience with and knowledge about the environment, and we conclude that the robot must have learned something about its environment too.

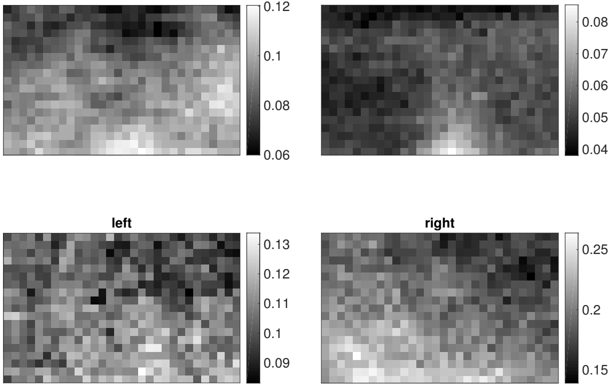

Fig. 4. MI of the motion flow and the utility in the subsequent epoch for the robot's four movement directions.

We propose that the MI maps in Figure 4 can be used as attention maps that restrict processing of visual information to the image regions that contain relevant information about the task to solve. Applying this idea to our collision avoidance scenario suggests that during forward and backward movements, only the bottom center and the lower side-lobes in the visual field need to be evaluated, whereas during left- and rightward movements, visual processing can be restricted to regions on the lower border and opposite side to the movement direction of the vertical axis. This highly adaptive processing schedule could contribute to a more efficient resource utilization in autonomous robots.

Using MI for attention maps is reminiscent of the use of saliency maps for restricting regions of visual search [12], [13]. Whereas these saliency maps analyze each image completely in order to highlight interesting regions, our MI attention maps are activated by the agent's action context, i.e. before starting to analyze the image data. This difference is an epitome for the distinction between bottom-up and top-down attention. During the learning phase, however, our approach works bottom-up as well. Complete images have to be analyzed in order to learn the SMCs of motion flow and compute the informational value of regions in the visual field across different actions.

Another way of utilizing these attention maps is to weigh the contribution of each image location, respective to its relevance, in the decision process. Theoretic models for optimizing information intake in regard to a specific utility function have been proposed in [19], and their adaptation to this particular scenario is one potential direction for future research.

Even if modern bottom-up attention approaches show improved performance by considering various high-level constraints, they still do not generalize well to viewing behaviors in natural environments. The main problem is that such action-agnostic models may adequately describe how static scenes are viewed, but that this does not carry over to analyzing image streams in dynamic explorations of natural environments [20]. Very likely such dynamic, action-based vision requires other mechanisms than processing snapshots sequentially and independently. A good example of how action-based attention mechanisms may cope with real-world tasks is given in [21], where an agent learns different policies for overt gaze shifts in different micro-behaviors (litter collection, sidewalk navigation and collision detection) through reinforcement learning. The results we obtained here suggest that this should be feasible also in our scenario, where gaze shifts would be covert and the collision status would serve as the reward function. Demonstrating that the attentional mechanism we propose here can help control the agent's behavior will be the focus of our future studies.

REFERENCES

[1] J. O'Regan and A. Noë, "A sensorimotor account of vision and visual consciousness," *Behavioral and Brain Sciences*, vol. 24, pp. 939–1031, 2001.

[2] A. Maye and A. K. Engel, "Context-dependent dynamic weighting of information from multiple sensory modalities," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 2812–2818.

[3] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Computing Surveys (CSUR)*, vol. 27, no. 3, pp. 433–466, 1995.

[4] R. C. Nelson and J. Aloimonos, "Obstacle avoidance using flow field divergence," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 11, no. 10, pp. 1102–1106, 1989.

[5] D. Coombs, M. Herman, T.-H. Hong, and M. Nashman, "Real-time obstacle avoidance using central flow divergence, and peripheral flow," *Robotics and Automation, IEEE Transactions on*, vol. 14, no. 1, pp. 49–59, 1998.

[6] A. Branca, E. Stella, and A. Distante, "Passive navigation using egomotion estimates," *Image and vision computing*, vol. 18, no. 10, pp. 833–841, 2000.

[7] F. G. Meyer, "Time-to-collision from first-order models of the motion field: Perception-based real-world navigation," *IEEE transactions on robotics and automation*, vol. 10, no. 6, pp. 792–798, 1994.

[8] J. M. Galbraith, G. T. Kenyon, and R. W. Ziolkowski, "Time-to-collision estimation from motion based on primate visual processing," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1279–1291, 2005.

[9] W. Reichardt and W. Rosenblith, "Autocorrelation, a principle for evaluation of sensory information by the central nervous system," in *Symposium on Principles of Sensory Communication 1959*. MIT Press, 1961, pp. 303–317.

[10] M. Egelhaaf, R. Kern, and J. Lindemann, "Motion as a source of environmental information: a fresh view on biological motion computation by insect brains," *Front. Neural Circuits*, vol. 8, no. 127, 2014.

[11] N. M. Schmidt, M. Hoffmann, K. Nakajima, and R. Pfeifer, "Bootstrapping Perception using Information Theory: Case Studies in a quadruped Robot Running on Different grounds." *Advances in Complex Systems*, vol. 16, no. 2-3, 2013.

[12] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[13] L. Itti and C. Koch, "Computational Modeling of Visual Attention," *Nature Reviews Neuroscience*, vol. 2, pp. 194–203, 2001.

[14] A. Maye and A. Engel, "Time scales of sensorimotor contingencies," in *Brain-Inspired Cognitive Systems 2012 (BICS 2012)*, ser. LNAI, H. Zhang, Ed., vol. 7366. Springer-Verlag Berlin, Heidelberg, 2012, pp. 240–249.

[15] A. Maye and A. K. Engel, "Using sensorimotor contingencies for prediction and action planning," in *From Animals to Animats 12*, ser. LNCS, T. Ziemke, C. Balkenius, and J. Hallam, Eds., vol. 7426. Berlin Heidelberg: Springer, 2012, pp. 106–116.

[16] T. Schreiber, "Measuring Information Transfer," *Phys. Rev. Lett.*, vol. 85, no. 2, pp. 461–464, 2000.

[17] J. T. Lizier and M. Prokopenko, "Differentiating information transfer and causal effect," *Eur. Phys. J. B*, vol. 73, pp. 605–615, 2010.

[18] N. Ay and D. Polani, "Information Flows in Causal Networks," *Advances in Complex Systems*, vol. 11, no. 1, pp. 17–41, 2008.

[19] D. Polani, C. Nehaniv, T. Martinetz, and J. T. Kim, "Relevant Information in Optimized Persistence vs. Progeny Strategies," *Proc. Artificial Life*, 2006.

[20] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard, "Eye guidance in natural vision: Reinterpreting salience," *Journal of vision*, vol. 11, no. 5, p. 5, 2011.

[21] N. Sprague, D. Ballard, and A. Robinson, "Modeling Embodied Visual Behaviors," *ACM Trans. Appl. Percept.*, vol. 4, no. 2, July 2007.