Stereo Matching and Depth Perception by Visual Prediction

Alexander Kaiser[1], Wolfram Schenck[1,2], and Ralf Möller[1,2]

[1] Computer Engineering Group, Faculty of Technology, Bielefeld University
[2] Center of Excellence Cognitive Interaction Technology (CITEC), Bielefeld University

1 Introduction

Recent theories of cognition state that perception is facilitated by the internal simulation of sensorimotor processes (see e.g. [1]). Internal simulation is the process of generating covert sensory states as the outcomes of covert motor commands, thus mentally simulating specific actions. In the case of perception, the internal simulation itself is considered as an unconscious process. In this article we present a computational model for stereo matching based on an internal simulation process. Moreover, the sequence of covert motor commands can be used to infer depth information from a pair of stereo images.

Covert sensory states and motor commands that arise in internal simulations will be referred to as mental images. This is in line with psychological findings that dichotomously classify mental images in visual [3] and motor imagery [2]. We are aware that classically mental imagery refers to conscious processes [3, 2]. In contrast to that, our use of the term is more general and also includes sensory states arising in unconscious internal simulations.

We apply the paradigm of internal simulation to the problem of stereo matching, i.e. the problem of finding correspondences in a pair of images depicting the same scene, albeit from slightly different view points. In contrast to classical matching algorithms that work on image information alone, our novel approach operates on the sensorimotor domain. Moreover, we work with retinal images that mimic the cone distribution on the human retina. These images have a higher resolution towards the center and a low resolution in the periphery. The resulting fish-eye effect makes it very difficult to find correspondences and most "classic" matching approaches are likely to fail in this situation.

We tested our model on an arrangement of simple objects located on a table in front of the cameras. A primitive attention mechanisms based on color segmentation is used to select potential targets. Once a target is selected in the left camera image, its corresponding match in the right camera image is computed using the internal sensorimotor simulation. The experiments suggest that our model is working reliably under controlled conditions.

2 Setup

The cameras used for our study are color cameras whose images are grabbed with a resolution of 320×240 pixels. The camera images where cropped and distorted by a radial foveal mapping, resulting in retinal images of size 207 × 207; these images were actually used as inputs to our model.

Each camera is mounted on a separate pan-tilt unit. Thus, both cameras can be moved independently. The PTUs have a pan range of $-60.4°$ to $23.8°$ and a tilt range of $-42.9°$ to $21.4°$. Each PTU is constructed such that its pan and tilt axes intersect in the vicinity of the entrance pupil of the camera. Thus, the depth of the scene is irrelevant and has no effect on the change of the visual content of the images when the PTUs are moved.

3 Computational Model

The computational model for stereo matching mainly consists of two internal models, a saccade controller (inverse model) and a visual forward model, whose interplay will be described in the following. Furthermore, the peripheral units for object selection and correlation-based matching will be described briefly.

3.1 Saccade Controller

The saccade controller (SC) [4] is an inverse internal model that generates a fixation saccade (motor command) $(\Delta\rho_1, \Delta\theta_1)$ given a fixation point (x_0, y_0) in retinal image coordinates and the current tilt angle θ_0 of the PTU.[3] The generated motor command is not executed, but sent to the visual forward model that is used to predict the position of the initial fixation point (x_0, y_0) after the saccade, denoted by (x_1, y_1). This new position, along with the updated tilt angle $\theta_1 = \theta_0 + \Delta\theta_1$, is fed once more into the SC in order to refine its initial estimate. This procedure is repeated N times, leading to a sequence that is accumulated to yield the final fixation saccade: $\Delta\rho_f = \Delta\rho_1 + \cdots + \Delta\rho_N$ and $\Delta\theta_f = \Delta\theta_1 + \cdots + \Delta\theta_N$. For our experiments, we used $N = 5$ correction steps.

3.2 Visual Forward Model

The visual forward model (VFM) [5] predicts an image $\hat{I}$ based on a (current) input image I and a given saccade $(\Delta\rho, \Delta\theta)$. The predicted image appears as if the saccade would have been executed by the PTU. Using this model, only portions of the image already present in the input can be predicted; the other portions are marked as invalid by the VFM and their pixel values are set to zero.

For its application to the task of stereo matching, the visual forward is used to bring the objects to be matched in a canonical—i.e. fixated—view. The necessary motor command is generated by the SC. At first, the selected object in the left (dominant) retinal image is fixated by the VFM, yielding a "mental" image of the object. This mental image is then compared to the fixated candidates from the right (subordinate) retinal image, by likewise applying the SC / VFM. With both, selected object and potential partners fixated, simple matching algorithms can be applied.

[3] Supplying the current tilt angle as an additional input was necessary in order to improve the performance of the SC. Still, the SC did not achieve the desired precision. Therefore, a sequence of correction saccades needed to be generated.

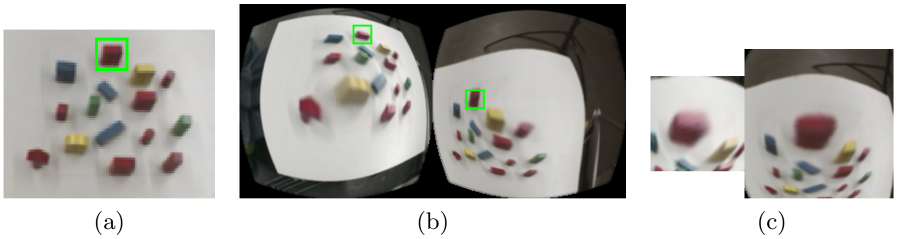
(a) (b) (c)

Fig. 1: Planar view of the scene (a), left and right retinal images (b), mental images of the fixated salient object (left) and the best match (right) (c).

3.3 Image Processing and Matching

After grabbing, the raw left and right camera images are processed by the radial foveal mapping. The resulting retinal images are segmented according to a specific salient color. In our experiments, we used the color red. From the resulting list of segments, the biggest one (according to pixel count) in the left image is selected. The centroid of this segment is used as the input to the SC. Afterwards the VFM is used to predict the new fixated image based on the left camera image and the saccade. From this image, a window of 101 × 101 pixels around the center is cut out. The segment list of the right image is iterated and each segment is processed analogously to the left image. Finally, the cut-out containing the salient object is matched against each of the "mentally" fixated candidate objects using the angle cosine (i.e. the dot product divided by the product of the vector norms) between the two images (interpreted as vectors).[4] The candidate yielding the highest value is recognized as the corresponding match.

4 Experiments and Results

We conducted a simple experiment for the stereo matching algorithm: an arrangement of objects made out of colored wooden blocks was placed in front of the camera-PTU setup. Figure 1a shows a planar views of the scene. All objects are colored uniformly. A simple attention mechanism based on a color segmentation selects one of these objects in the left retinal image. The same color segmentation algorithm is then applied to the right retinal image. The resulting segments are iterated and matched against the selected object by the internal simulation process (see Sec. 3).

Before the experiment, the PTUs oriented the cameras into random directions, such that the radial foveal mapping makes the images appear rather dissimilar (see Fig. 1b). An example salient object is framed by a green square in Figs. 1a and 1b. After mental fixation the salient object appears in the image center (Fig. 1c, left side). The fixated view of the best matching candidate is shown

[4] Formally, the angle cosine is defined by $\cos\beta = \frac{\boldsymbol{a}^\top \boldsymbol{b}}{\|a\|\|b\|}$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ denote the vectors containing the RGB-values of the two images to be matched. The value of the angle cosine lies within the range $[-1, 1]$.

in Fig. 1c, right side. Note that Fig. 1c shows the unprocessed output of the VFM. The corresponding match value is ~ 0.98. The experiment was repeated with all other objects in the left retinal image, yielding a success rate of 100%, i.e. all 8 salient objects were successfully matched against their corresponding partners in the right retinal image.

5 Conclusions and Outlook

We presented a biologically oriented computational model for stereo matching based on internal simulation. The covert sensory states and motor commands that occur during the simulation process can both be regarded as forms of mental images, respectively. Experiments show that our model can reliably match salient objects in a pair of stereo images. The main advantage of our approach is the inclusion of sensorimotor instead of purely sensory information. Thus, our model is invariant under the retinal representation with respect to the current gaze direction. Classical algorithms that consider sensory (visual) information alone are likely to fail in such a setting.

Furthermore, the model can be easily extended to extract depth information from the covert fixation movements generated during the internal simulation. As one can easily see, the vergence angle, defined by $\alpha = \rho^L - \rho^R$, where ρ^L, ρ^R denote the pan angles of the left and right PTU, is anti-proportional to the depth of the fixated object. This relationship could be used to assign a depth value to each object from the scene.

References

1. Hesslow, G.: Conscious thought as simulation of behaviour and perception. Trends in Cognitive Sciences 6(6), 242–247 (2002)
2. Jeannerod, M.: Mental imagery in the motor context. Neuropsychologia 33(11), 1419–1432 (1995)
3. Kosslyn, S.M., Alpert, N.M., Thompson, W.L., Maljkovic, V., Weise, S.B., Chabris, C.F., Hamilton, S.E., Rauch, S.L., Buonanno, F.S.: Visual mental imagery activates topographically organized visual cortex: PET investigations. Journal of Cognitive Neuroscience 5(3), 263–287 (1993)
4. Schenck, W., Möller, R.: Staged learning of saccadic eye movements with a robot camera head. In: Bowman, H., Labiouse, C. (eds.) Connectionist Models of Cognition and Perception II, pp. 82–91. World Scientific (2004)
5. Schenck, W., Möller, R.: Training and application of a visual forward model for a robot camera head. In: Butz, M.V., Sigaud, O., Pezzulo, G., Baldassarre, G. (eds.) ABiALS: From Brains to Individual and Social Behavior, pp. 153–169. No. 4520 in Lecture Notes in Artificial Intelligence, Springer (2007)