Continuous and robust saccade adaptation in a real-world environment

Tobias Rodemann, Frank Joublin, Christian Goerick

2006

Preprint:

This is an accepted article published in KI-Künstliche Intelligenz. The final authenticated version is available online at: https://doi.org/[DOI not available]

Published in KI - Künstliche Intelligenz, Vol. 03 , 2006, pages 23–26

Continuous and robust saccade adaptation in a real-world environment

Tobias Rodemann, Frank Joublin, Christian Goerick

To saccade with a camera system to an interesting visual stimulus requires a mapping from image coordinates (pixel positions) to motor coordinates (e.g. pan and tilt angles). Because this relation can depend on many parameters, some of which might be varying over time, it is advisable to learn this mapping. However, most of the existing solutions require a special calibration procedure or at least a cooperating environment. In this work we describe a system, inspired from the human oculo-motor system, that can continuously adapt the saccade control in a real-world environment. The focus is not on speed of adaptation or the precision of saccades but on the robustness of the adaptation process. Our system can operate in feature-rich environments, monitor its own performance, and is not impaired by external motion.

1 Introduction

Learning is one of the most fundamental aspects of biological systems and an area where technical devices are still inferior to their biological counterparts. For the field of robotics and any technical system which has to behave / act in real-world environments the relation between sensory inputs and motor output is of highest importance. A sensory-motor system that has been studied extensively due to its relative simplicity is the saccade control system [4]. The term saccade denotes feedforward controlled eye movements toward a (visually) salient target. The core problem of saccade control is to find the relation between the position of a salient stimulus on the retina (or the pixel position in a digitized image) and the corresponding gaze directions of the eye or a camera to focus on this object. The actual relation depends on a number of parameters (internal and external camera and head parameters). Using a special calibration procedure these parameters can be measured. The calibration procedure has to be repeated whenever any relevant part of the system has been changed. This can e.g. happen when the optics of the camera or the overall design of the robot's head is changed, or some software modules within the processing chain have been modified (for example a rescaling of image sizes). With increasing system complexity, more parameters have to be kept track of and the chances to decalibrate the system get higher. In addition to purposely made changes there is also the danger of mechanical wear-down, minor mechanical or electrical errors, and temperature sensitivity of some control elements. For biological systems the possible range of variations is even bigger, especially over the lifespan of an individual. The possibility and the potential of learning approaches for saccade control has been shown before, see e.g. [2, 7, 9]. The focus in these approaches is on speed of learning and precision, not on the robustness of the learning process, that is, the ability to learn in various uncooperative environments. The work of Bruske et al. [2] for example uses a special 2-stage calibration procedure with a single salient stimulus under direct control (a light source attached to a robot arm performing random movements). Nature can obviously live without these constraints, showing a high flexibility in natural environments. Experiments have shown [5] that humans and animals are perfectly capable of readapting their saccade control if their visual input is distorted, e.g. through prism lenses. This re-adaptation occurs during normal life in an every-day environment. It is this remarkable adaptability that motivated us for our research. In this work we present a system that can learn and readapt a saccade mapping for a 2-degree-of-freedom camera-head online in a real-world environment. The basic idea of the system is to compare pre- and post-saccadic inputs and find matching regions in the two images (see also [8]). By linking these regions to the executed motor command it is possible to learn the relation between motor coordinates and image coordinates. However, this simple approach is hampered by a number of factors related to working in a real-world environment:

1. External control of saccade targets. The control of saccades is not given to the adaptation algorithm (e.g. no random saccades), but rather determined by the objectives of the overall system, like scene exploration or object recognition. So adaptation has to occur in parallel to and independently of the control of the head.
2. Feature-rich environment. To find out where parts of an image moved due to a saccade, one has to solve the correspondence problem. This is far more difficult in a natural environment which contains many objects but also large homogeneous surfaces (e.g. walls).
3. External motion. A typical occurrence in real-world environments is motion of external objects leading to a change of the sensory input which is not related to saccadic motion.
4. Limited error tolerance. While saccade control is one of the less critical motor control actions, still it is advisable to make sure that the mapping, after being decalibrated, doesn't become worse through the adaptation, or, that the latter even causes a decalibration by itself.

Our approach can deal with these problems using a confidence measure (section 3.3), motion detection (section 3.4), and adaptation control (section 3.5).

2 Implementation framework

The saccade adaptation system is part of a larger project of an integrated system for active sensing and pattern recognition

[3], called BASS (brain-like active sensing system). The system consists of a neck and a camera system (see Fig. 2, left) that provides images, a preprocessing stage for saliency computation, an object identification system, modules for gaze target selection, modules for the memory trace, and the modules for online-adaptation of gaze control (Fig. 1). The neck has 2 degrees of freedom (pan and tilt angles) but the rest of the system is considered to be fixed. We are working only with the left camera for our saccade system. An important subtask for the system is the fovealization of objects with the target of object recognition. Saccade targets are chosen based on the computation of saliency maps for e.g. color, edges, or motion, while also using an inhibition-of return type mechanism [6] to avoid visiting the same locations again and again. This implies that the target of the gaze selection changes from one saccade to the next, which is another important constraint for the adaptation scheme. The position of salient objects has to be transformed into motor coodinates to fovealize the objects. This coordinate transformation is exactly the one that has to be learned for the saccade control - a mapping from 2D image coordinates to 2D motor coordinates.

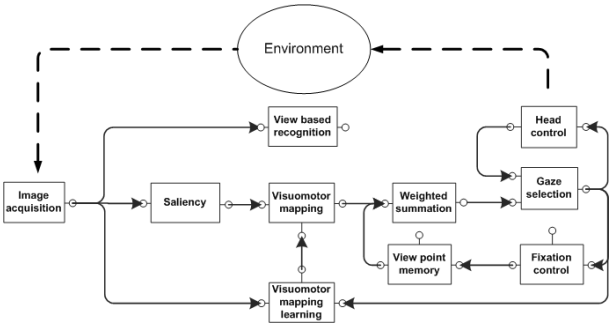

Abbildung 1: Schematic description of the BASS system. The interaction with the environment is done via sensory (*image acquisition*) and motor (*head control*) modules. From the images saliency maps are computed and transformed into motor coordinates by the *visuo-motor mapping* module. The latter contains the sensory-motor mapping that is adapted in the *visuomotor mapping learning* module. Based on saliency and *view point memory* a gaze target is selected.

3 Methods

The mapping $W(x,y)$ we have to learn is the correspondence between image point (x,y) (pixel coordinates) and motor coordinate $\vec{M} = (\phi, \theta)$ (corresponding to pan (ϕ) and tilt (θ) positions of the head motor element). The map entry $W(x,y)$ contains a motor command vector (ϕ, θ) that moves an object originally at position (x,y) into the fovea. The range of x, y is given by the size of the image (in pixels), the range of gaze directions by the chosen operation range of the head motor. We learn the sensory-motor relation by comparing changes in the position of objects induced by changes in the motor position of the head. We do not pick a specific object beforehand and look where it ends up after the saccade, but rather where the image patch that ends up in the fovea after the saccade was before the saccade. That means we identify the image position that was moved into the fovea by the issued motor command. Therefore we can directly associate sensory and motor coordinates without any need for inter- or extrapolation. This approach requires a feature-rich visual input as found in almost all natural environments. We now shortly describe the basic concept of the learning algorithm and then elaborate more on the details.

3.1 Image Correspondences

Our working environment contains a large number of objects, textured background, and various sources of illumination, see Fig. 2 (right panel). To find the correspondence between an image patch in the pre-saccadic image and the content of the fovea of the post-saccadic image we compute a correspondence function $C(x,y)$ between the fovea and pre-saccadic image patches at all positions (x,y). The correspondence between the post-saccadic foveal patch $\vec{r}^f$ and pre-saccadic image patch $\vec{r}(x,y)$ around image position (x,y) is given by:

$$C(x,y) = \frac{\vec{r}^f \cdot \vec{r}(x,y)}{||\vec{r}^f|| \cdot ||\vec{r}(x,y)||} \quad . \tag{1}$$

It turned out that performance is best if image patches are not normalized for brightness and contrast, which led to frequent mismatches in our tests (also note that there is no automatic white-balancing in our cameras). We used image patches of dimension 51 x 51 pixels (full image size 348 x 256 pixels). The point of maximum correspondence $(x_{max}, y_{max}) = \underset{x,y}{\text{argmax}}\, C(x,y)$ is then taken to be the position of the image patch that was moved by the executed motor command $\vec{M}_e$ to the fovea. We thus have to learn the association between $\vec{M}_e$ and (x_{max}, y_{max}).

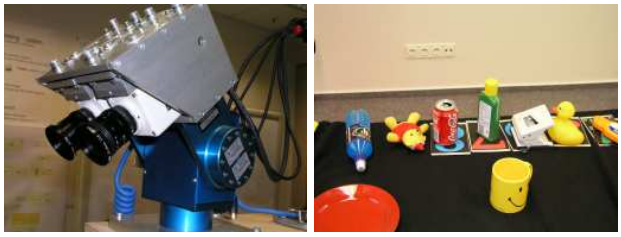

Abbildung 2: *Left:* The stereo cameras on a pan-tilt head element. *Right:* Example snap-shot of the visual environment used for testing our adaptation mechanism.

3.2 Adaptation Algorithm

After finding the corresponding image patches we update the connection matrix $W(x,y)$. For an executed motor-command $\vec{M}_e = (\phi_e, \theta_e)$ and correspondence position (x_{max}, y_{max}) we do the following adaptation step:

$$\Delta W(x,y) = -\alpha \cdot \kappa \cdot G(x,y,x_{max},y_{max})(W(x,y) - \vec{M}_e) \quad , \tag{2}$$

with α as the learning step size and κ as the confidence value (see section 3.3). To improve performance we adapt also vectors in the vicinity of the best matching one. We take a Gaussian neighborhood function $G(x,y,x_{max},y_{max}) =$

$\exp\left(-\frac{(x-x_{max})^2+(y-y_{max})^2}{\sigma^2}\right)$, which reduces the degree of adaptation with increasing distance from the best matching vector. In this equation σ is the width of the adaptation region (a system parameter). In section 3.5 we explain how to set σ dynamically.

3.3 Confidence Measure

One of the cornerstones of our approach is to assign a confidence value to each adaptation step, reducing the adaptation, or even canceling it altogether in case of a low confidence. The main source of errors for our approach is due to mismatches in the correspondence measurement process. Therefore we introduce the confidence measure κ. First we compute a sigmoidal of the maximum correspondence value c_{max} from eqn. 1:

$$\kappa' = \frac{1}{1+\exp(-c_s \cdot (c_{max} - c_t))} \quad . \tag{3}$$

This is a sigmoidal function with threshold value c_t and a slope c_s. We received good results for $c_s = 20$ and $c_t = 0.75$. This step reinforces correspondence values above c_t and suppresses those below. To reduce the confidence in case of multiple good matches (a common problem for homogeneous patches) we perform a normalization (division) of the confidence value by the number of entries in the correspondence map with a value above a threshold $T = 0.9 \cdot c_{max}$:

$$\kappa = \kappa' \cdot \frac{1}{\sum_{x,y} N(x,y)}, \quad N(x,y) = \begin{cases} 1 & : \quad C(x,y) > T \\ 0 & : \quad \text{else} \end{cases} \quad . \tag{4}$$

The confidence measure κ is used in equation 2 to modulate the adaptation step size.

3.4 External Motion

Apart from mismatches in the correspondence search there is another factor that quickly leads to errors in the adaptation process: external motion. Our BASS system is designed to interact with humans in a natural way and therefore the environment is not static but rather contains many moving objects. We observed that this strongly impairs the adaptation. Therefore we added a motion detection system that signals external motion. This system compares two consecutive camera images (using the same camera as for the adaptation) and when the difference exceeds a threshold value, the following adaptation step is canceled. Motion detected during the ego-motion of the camera is not taken into account.

3.5 Adapting the Adaptation

Two parameters control the adaptation process: the step-size of adaptation α and the width σ of the population that is adapted. For the step size α a constant value of 0.8 led to good results, i.e. a rapid adaptation as can be seen in Fig. 3. To adapt the population width σ we compute the saccade error E: the difference (in pixels) between the pre-saccadic position of the object that was moved to the fovea and the position that was set as the target by the gaze selection system. To be invariant to image size we divided the saccade by the maximum possible saccade length E_{max} (half image diagonal = 216 pixels). We have chosen this normalization instead of e.g. dividing by the intended saccade length because for our application the absolute saccade error is more important than the relative one, since we want to get the target patch into a fovea region. We use the mean saccade error $\bar{E}$ (E averaged over the last 10 saccades) as a measure to modify the adaptation width:

$$\sigma = \sigma^{max} \cdot \frac{1}{(1+\exp(-s \cdot (\bar{E} - t)))} \quad . \tag{5}$$

In this equation σ^{max} is the maximum adaptation width (a parameter that has to be set in relation to the size of the image), s the slope of the sigmoid, and t the threshold of the sigmoid. Increasing saccade errors $\bar{E}$ lead to an increased adaptation width σ. In our experiments we used $\sigma_{max} = 30\%$ of the size of the image, $s = 20$ and $t = 0.2$. The result is that the mapping will be very flexible when the saccades are far off target, but will be modified only locally when the performance is good. Learning was very robust over many trials and different initial conditions. However, allowing σ to get too high ($> 30\%$) made the learning process unstable.

4 Results

We evaluated our adaptation algorithm within the complete BASS system in the described real-world environment. As a test we let the system adapt from a random initialization. The saccades were chosen to explore the environment and track salient objects. Within less than 100 saccades the map was structured and the system showed correct saccade behavior. We then introduced a (software) prism effect that inverted the image on the horizontal axis (upside-down). The system quickly adapted without any user interference or parameter changes and also readapted when the prism was removed. Figure 3 shows an example run of our algorithm. The system performed the saccading and object recognition task in parallel to the learning, which was never disabled and worked without problems for many hours. The final accuracy of the system after a longer adaptation period (ca. 1000 saccades) is in the order of 10 pixels. This sometimes requires a second corrective saccade to fovealize a target, as it is also observed for the human eye [1]. This residual error is largely invariant of adaptation parameters (e.g. α) and due to the fact that for our system setup precision is inherently limited because we can't determine the exact 3D position of the target which would be necessary for a correct mapping. Furthermore we used a comparatively small motor map with a limited resolution (1 degree). However, for our purposes the achieved precision of the saccades was satisfying. We did another test to evaluate the impact of external motion on the learning process. Using a learned mapping, humans began interacting normally with the system. We tracked the mean saccade error with and without the external motion detection system, see Fig. 4. It is obvious that undetected external motion can (temporarily) disrupt the adaptation process, when motion is unaccounted for. Using the motion detector to modulate adaptation, the saccade movements are largely unaffected by external motion.

5 Summary and Conclusion

We have presented a saccade adaptation in an integrated system for visual interaction in a real-world environment. The described

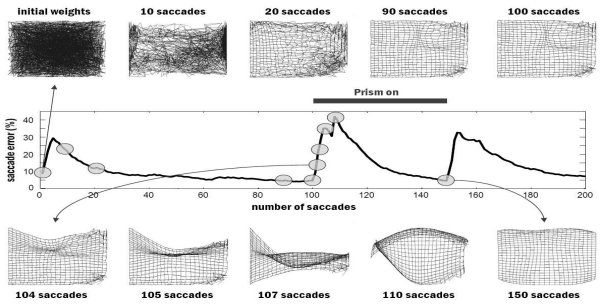

Abbildung 3: Mapping vectors for an example run of the saccade learning algorithm. Grid-points denote positions in motor space, while edges represent neighborhood relations in retinal space. Mapping vectors are initialized randomly and adapt to form a regular grid within a few 10 steps (top row). At step 100 an up-down inverting prism is added and the mapping quickly adapts to the new situation by flipping around (bottom row). The prism is removed at step 150 and the mapping flips again. The development of the mean saccade error is depicted in the middle row.

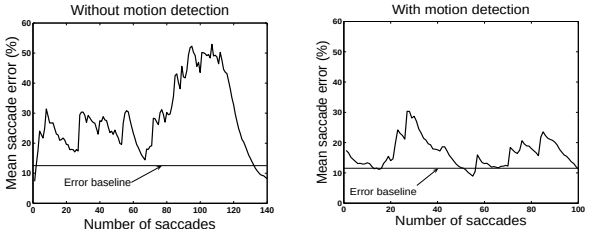

Abbildung 4: Development of the mean saccade error (in percent of $E_{\max}$ (216 pixels)) during system interaction with humans. The mean is computed over the last 10 saccades. The left panel shows the result without a motion detection system, the right panel with a motion detector. It is obvious, that the latter makes the adaptation process more robust. Note that the saccade error is generally higher during interaction with a user as saccades tend to be larger and more diverse in terms of target positions. With a disabled saccade adaptation the mean saccade error is around 12% in this scenario (*Error baseline* in figure).

system shows a reasonably quick learning with a satisfying level of precision. In contrast to many competing approaches we didn't focus on speed of adaptation or precision of the mapping but on the robustness of the process and its applicability for real online-adaptation. All required computations can easily run in real-time on standard PC hardware (less than 100 ms per saccade). The advantage of our approach compared to standard offline calibration procedures is that the system is usable without manual recalibration in case of hardware modifications (e.g. maintenance), smaller mechanical damages or software modifications. This can save lots of time and also makes special calibration set-ups unnecessary. This is specifically important for the field of robotics where calibration approaches require a deactivation of the robot. Our proposed solution allows a recalibration during operation.

Literatur

[1] Charles J. Bruce and Harriet R. Friedman. *Encyclopedia of the Human Brain*, volume 2, chapter Eye Movements, pages 269–297. 2002.

[2] Jörg Bruske, Michael Hansen, Lars Riehn, and Gerald Sommer. Adaptive saccade control of a binocular head with dynamical cell structures. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, and B. Sendhoff, editors, *Proceedings of ICANN'96, Lectures in Computer Science 1112*, pages 215–20, 1996.

[3] Christian Goerick, Heiko Wersing, Inna Mikhailova, and Mark Dunn. Peripersonal space and object recognition for humanoids. In *Proceedings of the IEEE/RSJ International Conference on Humanoid Robots (Humanois 2005), Tsukuba, Japan*, 2005.

[4] J. Johanna Hopp and Albert F. Fuchs. Investigating the site of human saccadic adaptation with express and targeting saccades. *Exp Brain Res*, 144(4):538–548, 2002.

[5] J. Johanna Hopp and Albert F. Fuchs. The characteristics and neural substrate of saccadic eye movement plasticity. *Progress in Neurobiology*, 72:27–53, 2004.

[6] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, 2000.

[7] Mitsuo Kawato. Feedback-error-learning neural network for supervised motor learning. In R. Eckmiller, editor, *Advanced Neural Computers*, pages 365–372. Elsevier, 1990.

[8] Michael Kuperstein. Neural model of adaptive hand–eye coordination for single postures. *Science*, 239:1308–1311, 1988.

[9] T. Shibata, S. Vijayakumar, J. Conradt, and S. Schaal. Biomimetic oculomotor control. *Adaptive Behavior*, 9(3/4):189–208, 2001.

Contact

Dr. rer. nat. Tobias Rodemann
Tel. +49 (0)69 89011-732
Email: Tobias.Rodemann@honda-ri.de
Dr. Ing. Frank Joublin
Tel. +49 (0)69 89011-727
Email: Frank.Joublin@honda-ri.de
Dr. Ing. Christian Goerick
Tel. +49 (0)69 89011-742
Email: Christian.Goerick@honda-ri.de

Honda Research Institute Europe
Carl-Legien-Str. 30
63073 Offenbach/Main
http://www.honda-ri.de

Bild

The authors are scientists at the Honda Research Institute Europe. The target of the institute is to understand and create brain-like intelligent systems based on the human brain and the biological evolution. A major research focus is learning and adaptation in real-world environments.