The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

A Neuro-dynamic Object Recognition Architecture Enhanced by Foveal Vision and a Gaze Control Mechanism

Christian Faubel and Stephan K. U. Zibner

Abstract— **We present an extension of a neuro-dynamic object recognition system that combines bottom-up recognition of matching patterns and top-down estimation of pose parameters in a recurrent loop. It is extended by an active foveal vision system. Adding the active vision component is easily integrated within the architecture and improves the recognition rate on previous experiments on the COIL-100 database and for scenes where segmentation of objects is not trivial. Furthermore the active component allows to substantially increase the spatial area where objects can be tracked. When objects move faster than visual servoing can track, catch-up saccades are autonomously generated.**

I. Introduction

Object recognition with autonomous robots differs substantially from the general object recognition problem in computer vision. The latter is still largely unsolved, especially when objects are embedded in natural environments [1]. The differences are rooted in the initial goal of recognition. Object recognition with an autonomous robot aims at enabling a robot to act autonomously, while in the general case of recognition in, for example, large image databases the aim is often to categorize in order to match a search pattern. In concrete robotic scenarios it is possible to define constraints on the problem of recognition that allow to simplify the problem compared to general object recognition and still fulfill the aim of enabling autonomous action. Here we consider a human-machine interaction scenario with our service robot CoRA [2]. CoRA is a stationary robot equipped with a seven degree of freedom arm and a pan-tilt stereo camera head. (see Figure 1). Objects relevant for interaction are those objects placed within the shared workspace and those which the robot can handle with its own gripper. The table itself being a homogeneous background, clutter is only introduced by other objects on the table. Occlusions may be induced by other objects, or the robot's own arm movement. Variations in scale between different objects are limited in so far as all objects have to fit into the gripper and have to be big enough to be graspable at all. Scale effects due to different distances to the imaging device are also limited because of the limited size of the interaction workspace. Objects have a typical standing axis, they can be rotated around this axis but full three-dimensional depth rotation will not occur. We have developed a neuro-dynamic architecture for one-shot learning of objects for such an interaction scenario. The system combines object recognition and object pose estimation in a recurrent loop of top-down prediction and bottom-up recognition [3].

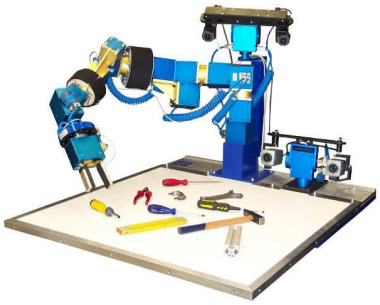

Fig. 1. The Cooperative Robotic Assistant, CoRA, with the shared workspace and items relevant for object recognition in front of it.

A major difference to feed-forward models [4], [5], [6] but also SIFT based approaches [7] to object recognition is the higher degree of autonomy of our system. For example recognition in feed-forward models requires a classifier at the output of the feed-forward feature processing. How such classifiers behave when being linked to the real-time output for example in the case of being connected to a continuous stream of video data from cameras is usually not discussed.

Here we make use of the higher degree of autonomy of our system and extend it by foveal vision and a gaze control mechanism. We can plug such a mechanism to the pose estimation module of our system without the need for any modification on the recognition and estimation system itself, because the object recognition system is endowed with stabilized representation and tracking capability [3]. Generally this is a non-trivial operation especially because a moving gaze adds a lot of variance to the camera images.

The initial motivation to switch to foveal vision was to speed up recognition by reducing input image size without loss of precision at the center. It turns out that not only the input image size was reduced by 20 per cent but also recognition performance improved compared to our previous effort on the COIL-100 database and on images where segmentation is non-trivial. In robotics a number of foveal vision systems exist. Foveal vision can be used to enhance segmentation [8], [9], in both papers the segmentation happens in the log-polar space and independent of recognition which is not addressed. Foveal vision has also been used in object recognition systems [10], [11]. While in the first

This work was supported by the German Federal Ministry of Research and Education (BMBF) through the the National Network Computational Neuroscience - Bernstein Fokus: "Learning behavioral models: From human experiment to technical assistance", grant FKZ 01GQ0951.

Christian Faubel and Stephan K. U. Zibner are with the Institut für Neuroinformatik, Ruhr-Universität Bochum, 47800 Bochum, Germany christian.faubel@ini.rub.de, stephan.zibner@ini.rub.de

paper the number of objects is limited to five, but real-time performance is demonstrated, in the second paper the number of objects is increased to hundred objects and recognition is demonstrated with eight learning views per object, but the real-time behavior of the system is not discussed.

When using foveal images a gaze control mechanism is needed, because in the extra-foveal areas the image is so distorted, that object recognition is severely affected. We use the estimate from the shift estimation module, which can also be interpreted as an attentional mechanism, to provide the control signal for gaze control. This is in analogy to findings in human and primate gaze control [12]. Our gaze control implementation combines saccadic eye movements and visual servoing by simulating foveal movement within the complete camera image combined with real head movement of CoRA's pan and tilt stereo head. The gaze control mechanism is largely based on a model of saccadic eye movement [13], [14], [15]. Most robotic systems that make use of active vision mechanisms rely on saliency-based attention models [16]. A recent example for such a system is presented by Dankers and colleagues [17]. In their paper the authors do not focus on object recognition but on generating human-like scan paths with the robotic eyes for tracking objects.

Because the fovea can move within the image and with the additional head movement, the range in which foveated objects may be tracked is enormously extended compared to our previous effort. With these additional degrees of freedom the system becomes a real visual tracking system that is robust to distractors because it uses a high-level object description that is also used to discriminate 30 objects. When object-based tracking is demonstrated this is usually done for a single object or objects of the same category such as people trackers [18], but not for larger numbers of different objects, see [19] for a survey.

II. The recurrent architecture for one-shot object learning.

The architecture we build on achieves recognition based on a small number of views [3]. Identifying an object from a small sample of such views is an inherently ill-posed problem, more so if the object's pose is to be estimated at the same time. Past efforts to address this problem have often taken inspiration from how the human nervous system seems to effortlessly solve the problem. Multiple feature histogramming approaches [20], [21], [4] for example have generated a degree of shift invariance through spatial pooling of feature representations and by learning the feature contributions that are most invariant for an object with respect to the possible remaining transformations. In order to uncover the invariant features, these approaches however require a larger number of training examples. Feed-forward view-based approaches also achieve shift invariance through a hierarchy of pooling stages [5], [6]. Invariance to rotation is only achieved by increasing the number of training views. An approach to limit the number of different views is to actively estimate the transformation an object has undergone relative to the learned view and to thus place the current view into an object-centered reference frame [22], [23]. The difficulty of such correspondence-based approaches is, of course, to uncover the transformation the object has undergone, which is far from trivial. Both estimation and pattern match must be computed in parallel. The recognition system is inspired by a recurrent architecture proposed by Arathorn, which solves this problem efficiently through pattern superposition and the cascading of multiple transformations [24].

Similar to this approach in a recurrent loop bottom-up information converges on a competitive dynamics that selects the recognized object while top-down information converges on a Dynamic Neural Field that estimates pose parameter values. The bottom-up path is based on feature channels, similar to multiple feature histogramming approaches [20], [21], [4]. Features are computed through pooling, both by summing over receptive fields to sample histograms [4] and by max-pooling operations [25] to generate shape templates. In parallel the top-down path computes estimates of the transformations between the current and learned representations of an object. Translational and rotational transformations are cascaded as in Arathorn's system [24].

At the core of the estimation process Dynamic Neural Fields [26] are at work. Conceptually, Dynamic Neural Fields are dynamical neuronal networks, in which the discrete sampling of relevant perceptual or motor dimensions by individual neurons is replaced by continuous distributions of neuronal activation (for a conceptual introduction, see [27]). Localized peaks of activation are units of representation. When the activation level in the peaks exceeds a threshold (conventionally chosen to be zero), such peaks represent perceptual or motor decisions, both in the sense of detection and in the sense of selection among competing inputs. The location of such peaks along the feature or motor dimension represents a metric estimate of the perceptual or motor state. Here the Dynamic Neural Fields enable the object recognition system to update its pose estimations online, so that the system can be coupled to the video-stream of images captured by the robot to track recognized objects. We use discrete dynamical neurons for the competitive selection of the winning memory pattern during matching. All modules are set as two layers of Dynamic Neural Fields or discrete neurons. The outputs of the estimation modules are used to modulate the superposition of extracted features in the bottom-up path.

The Shift Estimation Module

Because the shift estimation module is used to guide the eye movement of the foveal vision system we shortly review it here. For the details of the other rotation and pattern matching module and the details of the recurrent computation please refer to our previous work [3]. The shift estimation module consists of two Dynamic Neural Fields that are hierarchically organized. The field in the first layer (see Figure 2) directly receives input $s_s(x, y, t)$ from spatial feature correlation measures. The evolution of its activation variable $u_{s,1}$ defined over retinal space (x, y) in time t is

captured by the following dynamical equation:

$$\tau_1 \dot{u}_{s,1}(x,y,t) = -u_{s,1}(x,y,t) + h_1 + s_s(x,y,t) + \int\int w_{s,1}(x-x',y-y')\sigma(u_{s,1}(x',y',t))dx'dy' \quad (1)$$

Without input the field relaxes to its resting level defined by h_1, with input the field may locally pass threshold and build a peak defined through the field interaction which is expressed by the interaction kernel $w_{s,1}(x-x',y-y')$ and the threshold function $\sigma(u_{s,1}(x',y',t))$. It feeds its output to a second field $u_{s,2}$ that evolves on a slightly slower timescale $\tau_2 > \tau_1$ and is more interaction-dominated. Because it only receives input from the first layer and because of its slower timescale peaks only build after peaks have formed in the first layer. The first layer has broader and weaker interaction, multiple peaks are possible, whereas the second layer is strongly interaction-dominated and produces only single peaks that are spatially sharpened. Through the coupling of the two layers in the beginning of the recurrent computation the broader estimates are used to modulate the weighted sum of features in the bottom-up pathway and only later when a peak has built up in the second layer this spatially more precise estimate is used for computing the weighted sum. This modulation can be interpreted as an attentional mechanism that gates only the attended part of the input image.

The spatial correlations that are used as input are computed on a discrete grid by calculating the scalar product of the weighted sum of memory patterns in the top-down pathway and the feature distributions extracted around the points of the grid. These spatial computations are done for each feature channel and the resulting spatial maps are summed up with weight factors that were heuristically determined.

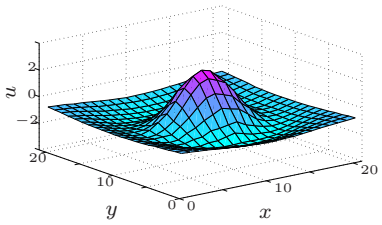

Fig. 2. *The first layer spatial shift field.* This plot shows the activation of the shift estimation field for an object that is centered on the image. A suprathreshold peak represents the spatial position of the object.

III. Foveal Vision and Gaze Control

A. Emulating Foveal Vision

To emulate foveal vision we first transform the input into the log-polar representation and then transform it back into the Cartesian representation as shown in Equation 2, but with a greater magnification factor $m_\mathrm{b} \geq m_\mathrm{f}$. Through this operation we obtain a smaller downsampled image that has its highest resolution at the pole (x_c, y_c) of the transformation and a smaller resolution at its borders (see Figure 3).

$$\begin{aligned} \rho &= m_\mathrm{f} log(\sqrt{(x_c-x)^2+(y_c-y)^2}) \quad (2) \\ \phi &= atan(y/x) \\ x &= x_c + e^{(\frac{\rho}{m_\mathrm{b}})} cos(\phi) \\ y &= y_c + e^{(\frac{\rho}{m_\mathrm{b}})} sin(\phi) \end{aligned}$$

Fig. 3. *Emulated foveal vision*. The left image shows the transformed input image. As one can clearly see the resolution at the center is higher. In the middle the representation in log-polar coordinates is displayed and on the right the input image is shown , the red square shows the region of interest that is transformed. The size of these images are 52×52 for the foveal image, 164×164 for the log-polar image and 219×219 for the region of interest.

This emulated foveal image has a smaller field of view than the original image and an increased variance in appearance to shift operations as one can see in Figures 3, 4 and 5. Shifting the pole of the transformation (x_c, y_c) moves around the foveated image and effectively simulates an eye movement in the original image. We use the spatial shift estimation to move the pole so that it is centered on the objects, and refer to this motion as simulated eye movement.

B. Simulated Eye Movement

For planning the eye movement we use a Dynamic Neural Field model of saccadic eye movement [14]. In this model a two-dimensional Dynamic Neural Field $u_m(x,y,t)$, the *motor planning field*, represents a motor plan for executing a saccadic eye movement. The *motor planning field* spans the possible movement directions in retinal coordinates (within the foveal image) and is thus defined over the same dimensions as the spatial shift estimation field. A peak centered around the foveal center specifies continued fixation, small movements of the input in this area lead to visual servoing behavior. A peak at extra-foveal locations represents a motor decision for triggering a saccade and the position of the peak encodes the corresponding distance to the target. Through global inhibition the field is set up for single peaks only, it will thus always perform a selection decision if presented with competing inputs. The switch between visual servoing mode and the saccade mode is realized by two competing discrete neurons $u_\mathrm{fix}, u_\mathrm{sac}$. One, u_fix, receives input that is integrated over the central-foveal area of the *motor planning field* and represents the fixation mode. The second, u_sac, receives input integrated over the extra-foveal area and represents the saccade mode. Both neurons are in competition through mutual inhibition, so that only one may be active at a time.

1) Saccadic Eye Movement: A saccade is a ballistic movement, the top velocity is proportional to the length of the movement, the duration of the saccade is fixed. To model the time course and the velocity profile we use a Hopf oscillator. The oscillator runs a single cycle with a duration that is determined by ω. It is switched on when the neuron representing the decision to initiate a saccade becomes supra-threshold ($\sigma(u_{\mathrm{sac}}) > 0$). This mechanism is represented by the first part in Equations 3 and 4. When the system is fixating, the competing neuron u_{fix} will be supra-threshold. In this case the oscillator is reset to its default state, which is $G_x = 1$ and $G_y = 0$. When the oscillator starts a saccade the *motor planning field* is boosted homogeneously. A working memory peak then self-stabilizes at the location of the saccade target.

$$\begin{aligned}\dot{G_x} = & \quad \frac{1}{\tau_{\mathrm{hopf}}}\left[\sigma(u_{\mathrm{sac}})(G_x - (G_x^2 + G_y^2)G_x - \omega G_y)\right] \\ - & \quad \frac{1}{\tau_{\mathrm{linear}}}\left[\sigma(u_{\mathrm{fix}})(G_x - 1)\right] \end{aligned} \tag{3}$$

$$\begin{aligned}\dot{G_y} = & \quad \frac{1}{\tau_{\mathrm{hopf}}}\left[\sigma(u_{\mathrm{sac}})(G_y - (G_x^2 + G_y^2)G_y + \omega G_x)\right] \\ - & \quad \frac{1}{\tau_{\mathrm{linear}}}\left[\sigma(u_{\mathrm{fix}})G_y\right] \end{aligned} \tag{4}$$

A saccade ends when the Hopf oscillator has gone through a single cycle. This is detected by a single variable that measures the distance of the Hopf variables G_x, G_y from their final position $s_{\mathrm{reset}} = \sigma(G_x - 0.7) \cdot \sigma(G_y - 0.2)$. This signal is used to shortly shut off all field activity through a homogeneous negative boost. When the field recovers from this boost, and if the saccadic eye movement has landed at the right position a peak induced by the pre-saccade target builds at the foveal center and the system switches back into fixation mode. If the saccade does not land at the right spot a peak will build in an extra-foveal area and a correction saccade is triggered.

The velocity is determined by the position of the Hopf oscillator in its cycle, maximum speed is reached at half of the cycle. The distance to travel is specified by the peak in the *motor planning field*. The two-dimensional *motor planning field* projects its output onto two one-dimensional fields $u_{\mathrm{hor}}(x,t)$ and $u_{\mathrm{ver}}(y,t)$ at a higher spatial resolution representing horizontal and vertical motion separately. The peaks in these fields represent the length of the saccade to execute in vertical and in horizontal direction. The peak positions can be approximated by treating the fields' activation as a probability distribution with means $\bar{x}$ and $\bar{y}$ at the peak positions.

2) Visual Servoing: When the system fixates it is not static but it is visually tracking the object through a visual servoing mechanism with strength α. Small displacements of objects lead to a peak that moves a little bit out of center but not into the extra-foveal region. Again we use the projection to the two one-dimensional fields $u_{\mathrm{hor}}(x,t)$ and $u_{\mathrm{ver}}(y,t)$, compute each mean, $\bar{x}$ and $\bar{y}$, and use these as servoing signals. Both the saccadic movement and the visual servoing are integrated into one single equation that computes the rate of change for each coordinate of the foveal transformation's pole. This equation is the same for both dimensions and here we only show it for the horizontal component.

$$\begin{aligned}\dot{x} &= \sigma(u_{\mathrm{saccade}})\bar{x}(1 - G_x)\frac{\omega}{2\pi} \\ &+ \sigma(u_{\mathrm{fix}})\alpha_{\mathrm{hor}}\bar{x} \end{aligned} \tag{5}$$

C. Head Movement

The simulated saccadic eye movements are limited to the field of view of the camera. In order to enable the system to attend the whole workspace we couple head movement that is performed with CoRA's pan-tilt unit to the eye movement. Based on the current position of the pole of the foveal transformation we compute a servoing signal for head movement that brings the pole to the center of the camera image. We use two one-dimensional Dynamic Neural Fields $u_{\mathrm{pan}}(\phi)$ and $u_{\mathrm{tilt}}(\theta)$ for representing these proprioceptive signals, which are defined over the corresponding opening angles of the field of view of the camera. Again from peaks in these one-dimensional fields we compute the mean values for $\bar{\phi}$ and $\bar{\theta}$. These are used to drive two identical simple dynamical systems of the pan and tilt angle of the head. Here we only show the equation for the pan angle.

$$\dot{\phi} = \beta\bar{\phi} \tag{6}$$

Speed of the head movement is adjusted by β. Both dynamical systems evolve on a much slower time scale than the simulated eye movement. The decoupling by timescales avoids oscillations of the whole motion system.

IV. Results

As a baseline test we first evaluate the recognition performance and compare it to the previous implementation. We use the same data-sets, our own data in which objects are placed at varied positions and orientations in the shared workspace of our robot CoRA and we test with thirty objects of the COIL-100 database.

In addition to these baseline tests we demonstrate the enhanced tracking capability of the new system in two experiments. The first experiment shows that recognition works on moving objects that move within the whole workspace and how the recognition then stabilizes tracking even in the presence of strong distractors namely other objects the system has already learned before. In the second experiment we show how the system autonomously switches from smooth pursuit to catch-up saccades during tracking, when the speed of the target object is gradually increased.

A. Recognition Performance

On our own data-set of thirty objects placed at nine different positions and orientations on the table, the system performs equally well and reaches a recognition rate of 90 per cent with only a single training view. Tested for two objects in a scene where segmentation is difficult, the new system reaches a recognition rate of 85 per cent on the first

autonomously chosen object. This is an improvement of 10 per cent points compared to our previous effort.

On the first thirty objects of the COIL-100 database the system achieves a recognition rate of 85 per cent with a single training image and tested on the remaining 71 views for each object. The previous system achieved this performance only with two training views. With two training views per object the system achieves a recognition rate of 96 per cent, which is what the previous system achieved with four training views.

B. Tracking

1) Tracking Performance and Distractor Robustness: In two experiments we demonstrate the tracking capability of the system and its robustness against distractors. To test the tracking we use a small mobile robotic platform that drives at a fixed velocity across the whole workspace. The mobile robot carries one of the objects, the *toothpaste*, an object the system has previously learned with a single shot. On its way the robot passes behind a distractor object, the *deodorant*, which is also known to the system and very similar in color. The system successfully tracks the moving object in a smooth pursuit mode and even sticks to it when it becomes partly occluded as can be seen in Figure 4. In a second experiment we invert the situation after the system has recognized the *deodorant*. The robot drives by with the toothpaste as distractor object. Despite this moving attractor the system sticks to the static object as is shown in Figure 5.

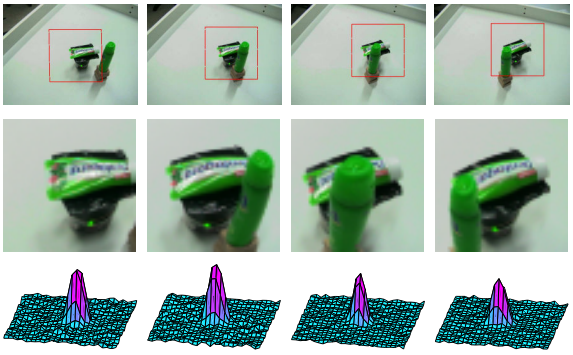

Fig. 4. *Moving object, static distractor.* In consecutive frames is shown the moving robot. The top images show the complete camera image. The current foveal area is marked with the square. Below is pictured the foveal representation the object recognition works on. The bottom row shows the field activation of the *motor planning field*.

2) Smooth Pursuit and Catch-up Saccades: To demonstrate that the system autonomously switches from smooth pursuit to catch-up saccades we gradually increased the robot's speed and let it run across the workspace. The system switches at a velocity of about 60 mm/s. Figure 6 shows snapshots from a smooth pursuit trial with the robot. When the robot moves faster, the peak still follows the robot but the servoing mechanism cannot catch up any more. The peak thus leaves the central foveal area which in turn triggers a saccade through the mechanism of the two competing neurons u_{fix} and u_{sac}. Because these saccades are triggered immediately when the peak reaches the extra-foveal area the saccades are very short. For reasons of illustration (Figure 7) we show thus a catch-up saccade where the object that is being tracked is moved with much higher acceleration than it is possible with the mobile robot.

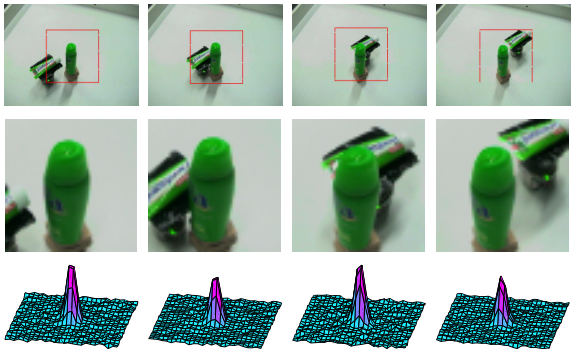

Fig. 5. *Static object, moving distractor.* The consecutive frames show the system fixating on the static object while the robot drives by. As can be seen in the last row any influence from the distractor object is suppressed in the *motor planning field*.

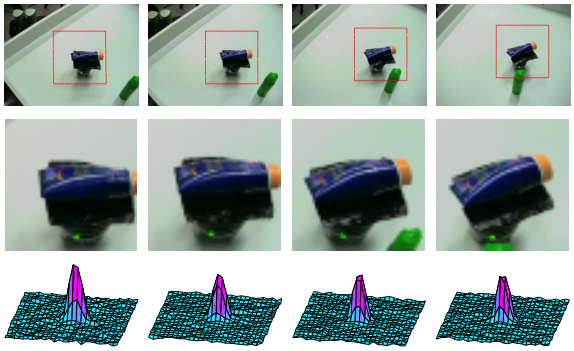

Fig. 6. *Smooth pursuit.* The figures show snapshots from the smooth pursuit experiment. The top images show the effect of the head movement that follows the eye movement, different parts of the workspace come into the field of view of the camera. The field activity plotted on the bottom row shows that the peak is always centered on the robot.

V. Discussion

We have enhanced an object recognition system by foveal vision and a gaze control mechanism. Not only the recognition performance is substantially improved compared to our previous effort, but also the system's tracking performance is greatly enhanced. Instead of the limited region of interest of the previous system now tracking is possible over the whole input image through the mechanism of virtual saccades. The area which can be tracked is further enhanced through the head movement so that effectively tracking is possible over the whole workspace. Both functionalities, recognition and tracking, interact smoothly in real-time as demonstrated in the tracking experiments.

The foveal representation leads to an increase in recognition performance in situations where objects are difficult to segment as in Figure 3 and on the COIL-100 database. The former is due to the fact that transformation into the foveal representations blows up the central part of the image and

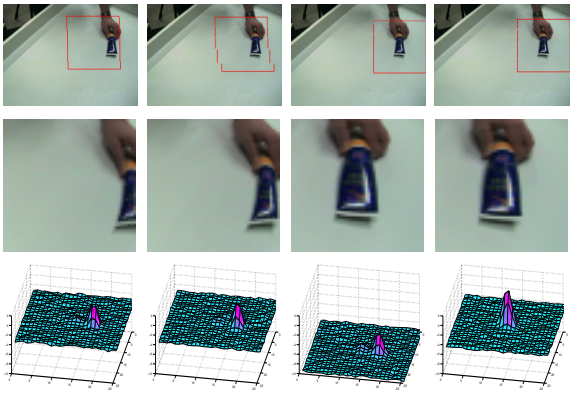

Fig. 7. *Catch-up saccades.* A human user moves the object so fast that the peak is not able to track and a new peak emerges at the new object location. This peak triggers a saccade. At the end of the saccade the field is de-boosted, as can be seen in the third plot from the left. When the field recovers a fixation peak builds up at the center of the fovea.

reduces the size of a distractor object as can be seen in Figure 3. The latter improvement on the COIL-100 database is because the log-polar transformation induces a higher degree of size or scale invariance. In the COIL-100 database objects are rotated and then scaled to fill the image, which induces a reasonable amount of size variance. When the images are transformed into the foveal representation pixels at the center are enhanced and extra-foveal pixels are diminished.

The seamless integration of a gaze control mechanism that introduces a high amount of motion that enters the feed-forward pathway of processing demonstrates the higher degree of autonomy compared to purely feed-forward based approaches to object recognition. The system is able to function under dynamic conditions. It stabilizes decisions, which then support the tracking behavior and all this happens without the need to explicitly set it into a different mode of functioning. Compared to most tracking approaches it can robustly track a higher number of different objects that can be learned with a single view.

The system is meant to provide object representation both of object identity and of pose parameters to a scene representation architecture that is currently under development [28]. Such a scene representation would for example allow to sequentially process objects that are in a scene, and a scene representation can provide context information that can be easily integrated into the dynamic architecture.

REFERENCES

[1] N. Pinto, D. D. Cox, and J. J. Dicarlo, "Why is real-world visual object recognition hard?" *PLoS Computational Biology*, vol. 4, no. 1, pp. e27+, January 2008. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.0040027

[2] I. Iossifidis, C. Theis, C. Grote, C. Faubel, and G. Schöner, "Anthropomorphism as a pervasive design concept for a robotic assistant," in *Proceedings of IROS 2003: The 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE Press, 2003.

[3] C. Faubel and G. Schöner, "A neuro-dynamic architecture for one shot learning of objects that uses both bottom-up recognition and top-down prediction," in *Proceedings of the IEEE/IRSJ International Conference on intelligent Robots and Systems IROS*, 2009.

[4] B. W. Mel, "Seemore: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition," *Neural Computation*, vol. 9, pp. 777–804, 1997.

[5] H. Wersing and E. Koerner, "Learning optimized features for hierarchical modelling of invariant object recognition," *Neural Computation*, 2003.

[6] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 994–1000. [Online]. Available: http://portal.acm.org/citation.cfm?id=1068508.1069194

[7] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. [Online]. Available: citeseer.ist.psu.edu/lowe04distinctive.html

[8] F. Orabona, G. Metta, and G. Sandini, "Object-based visual attention: a model for a behaving robot," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 3, 2005, p. 89.

[9] A. Mishra, Y. Aloimonos, and C. Fermuller, "Active segmentation for robotics," in *Proceedings of IROS 2009*, 2009.

[10] A. Ude, C. Gaskett, and G. Cheng, "Support vector machines and gabor kernels for object recognition on a humanoid with active foveated vision," in *Proceedings of the IEEE/IRSJ International Conference on intelligent Robots and Systems IROS*, 2004.

[11] K. Welke, E. Oztop, A. Ude, R. Dillmann, and G. Cheng, "Learning feature representations for an object recognition system," in *IEEE-RAS International Conference on Humanoid Robots - HUMANOIDS 2006*, 2006, pp. 290 – 295.

[12] W. Becker, *Metrics*. Elsevier, 1989, pp. 13–67.

[13] K. Kopecz and G. Schöner, "Saccadic motor planning by integrating visual information and pre-information on neural, dynamic fields," *Biological Cybernetics*, vol. 73, pp. 49–60, 1995.

[14] J. Goldberg, "When, Not Where a Dynamical Field Theory of Infant Gaze," Ph.D. dissertation, Indiana University, 2009.

[15] J. Goldberg and G. Schöner, "Understanding the distribution of infant attention: A dynamical systems approach," *Poster presented at the 29th Annual Cognitive Science Society. Nashville, TN*, 2007.

[16] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998. [Online]. Available: http://citeseer.ist.psu.edu/itti98model.html

[17] A. Dankers, N. Barnes, W. F. Bischof, and A. Zeliensky, "Humanoid vision resembles primate archetype," in *Experimental Robotics: The Eleventh International Symposium*, 2009.

[18] U. Handmann, T. Kalinke, C. Tzomakas, M. Werner, and W. Seelen, "An image processing system for driver assistance," *Image and Vision Computing*, vol. 18, no. 5, pp. 367–376, 2000.

[19] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, p. 13, 2006.

[20] M. Swain and D. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, pp. 11–32, 1991.

[21] B. Schiele and J. Crowley, "Object recognition using multidimensional receptive field histograms," 1996. [Online]. Available: citeseer.comp.nus.edu.sg/35791.html

[22] S. E. Palmer, "The psychology of perceptual organization: a transformational approach," *Human and machine vision*, pp. 269–339, 1983.

[23] B. A. Olshausen, C. H. Anderson, and D. C. Van Essen, "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information," *J. Neurosci.*, vol. 13, no. 11, pp. 4700–4719, November 1993.

[24] D. W. Arathorn, "Map-seeking circuits in visual cognition: A computational mechanism for biological and machine vision," Ph.D. dissertation, Stanford, CA, USA, 2002.

[25] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, pp. 1019–1025, 2000.

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

[27] G. Schöner, "Dynamical systems approaches to cognition," in *Cambridge Handbook of Computational Cognitive Modeling*, R. Sun, Ed. Cambridge, UK: Cambridge University Press, 2008, pp. 101–126.

[28] S. K. U. Zibner, C. Faubel, I. Iossifidis, and G. Schöner, "Scene representation for anthropomorphic robots: A dynamic neural field approach," in *ISR / ROBOTIK 2010*, Munich, Germany, 2010.