Learning to Anticipate the Movements of Intermittently Occluded Objects

Birger Johansson
birger.johansson@lucs.lu.se

Christian Balkenius
christian.balkenius@lucs.lu.se

Lund University Cognitive Science
Kungshuset, Lundagård
S-222 22 Lund, Sweden

Abstract

A model of event driven anticipatory learning is described and applied to a number of attention situations where one or several visual targets need to be tracked while being intermittently occluded. The model combines covert tracking of multiple targets with overt control of a single attention focus. The implemented system has been applied to both a simple scenario with a car that is occluded in a tunnel and a complex situation with six simulated robots that need to anticipate the movements of each other. The system is shown to learn very quickly to anticipate target movements. The performance is further increased when the simulated robots are allowed to cooperate in the tracking task.

1. Introduction

How do we know where a moving object will reappear after it disappears behind an occluder? Without any knowledge about the object, the best we can do is to assume that it will reappear at the same place where it disappeared. On the other hand, if the object was moving at constant velocity, we may assume that it will appear at the other end of the occluding object. More generally, if we know a little more about both the target object and the situation, we could form expectations of where and when the target object will reappear. It may also be useful to maintain a measure of how reliable these expectations are.

Piaget (1937) describes that a child is able to predict that a train that disappears at one end of a tunnel will appear at the other end. This can either be explained by a mechanism that continues to track the motion of the train after it has disappeared, or as a form of event learning where the child learns to predict that the disappearing train predicts the subsequent reappearance of the train.

Infants do not continue to track an occluded objects. Instead, one or two saccades are made to the other side of the occluder (Rosander and von Hofsten, 2004). These saccades are made to anticipate when the object reappears. This could be seen as an indication of an ability to predict the reappearance event based on the disappearance of the object. Wentworth and Haith (1998) found that three-month-old infants could learn spatiotemporal expectations of this type. In contrast, infants that are 7-9 weeks old continue to look at the edge of the occluder where the object disappears for 1 second before finding the target again (Rosander and von Hofsten, 2004). Infants that are 12 weeks old move their eyes as soon as the target becomes visible again. This delay decreases with each trial, which indicates that the infant starts to anticipate where the objects will reappear.

In 1958, Broadbent presented a single location attention mechanism that was separated from the visual fovea. This approach is suitable for single capability tracking but can not be used for multi target tracking. Instead other theories have been proposed for tracking of multiple targets. Either the attention system can use indexes, which switches a single attention focus between the different targets (Yantis, 1992; Pylyshyn and Storm, 1988), or it can use multiple parallel attention systems (Cavanagh and Alvarez, 2005). Humans are able to track 4-5 targets at the same time but this capacity vary substantially between individuals (Oksama and Hyoumlnauml, 2004). Pylyshyn and Storm (1988) used a multi object tracking set-up (MOT) to investigate our ability to track multiple targets moving at random order among distractors. Typically after 5-20 seconds the subjects are told to identify the targets depending on their initial positions. The number of targets that humans can track depends of the how the targets are presented. If targets and distractors are close to each other, the number of manageable targets decrease (Yantis, 1992).

Multi-tracking abilities are starting to appear in many different computer science applications. Systems using multiple cameras to track people, through

public places such as airports and underground stations and for surveillance purposes are widly used.

Our previous model of anticipatory gaze control consisted of three interacting paths from sensation to motor control (Balkenius and Johansson, 2007a). The first was a reactive saccade pathway that detects salient objects in the visual field and directs attention to them. The second part was an anticipatory pursuit pathway that learns to predict the motion of the target during the next fractions of a second. Finally, the event prediction pathway reacts to salient events and learn relations between them. This model was tested in simulations of the development of visual attention in infants and it was shown to parallel the developmental steps of gaze control during the first four months of life (Balkenius and Johansson, 2007a).

Here, we introduce two new components in the attention system. The first is a distinction between covert and overt attention, and the second is the ability to covertly track multiple targets. These two additions decouples the overt control of attention from the ability to track moving objects and makes it possible to maintain hypotheses about the location of several objects even when they are not visible.

The current work also extends our previous research by providing examples of how the event prediction system can handle real video input. In addition, we demonstrate how this system can be put to use in a complex cooperative multi-agent scenario with multiple moving targets and multiple visual obstacles.

2. Simple Anticipation

This section introduces the basic event prediction system within a tracking scenario and shows how it can be applied to video input to track an object that moves along a regular path.

2.1 The architecture of an event prediction system

We have used two forms of tracking methods with the event prediction system (Fig. 1). In the first case, we used a color based tracker (Balkenius and Johansson, 2007b). The localization of the target proceeds in three steps. The first is a color transform that converts the RGB image into a rgI representation consisting of a point in the rg-chromaticity plane together with intensity. In the next step, we classify each pixel in the image as being of the target color or not, and finally, a spatial clustering algorithm is used to group the pixels belonging to the target object. In the second case, we instead use a motion based tracker. This system detects any changes in the image which is subsequently clustered into a region that is assumed to contain the target.

Regardless of the visual processing method, the output sent to the event sensitive tracker consist of the coordinate of the target in the image and a value that codes how certain the target detection system is. A value of 1 indicates that the target is definitely present while a certainty value of 0 means that the target is not currently visible.

The tracking and event prediction is done by the module called AttentiveTracker in Fig. 1. This module uses the certainty value to determine whether the target is visible and detects two types of salient events. The first type is disappearance events that occurs when the certainty value falls below a threshold. The second type of event is an appearance event that occurs when the certainty rises above another threshold level. This is consistent with the idea that an event is any abrupt change in a variable (Prem et al., 2002).

When a target disappears, this is detected as an event E_1 and stored in working memory. This working memory contains both the location where the target disappeared $\langle x, y \rangle$ and the identity of the target. When the target later reappears this is considered a second event E_2 and the system learns an association between the two events. These associations are of the form,

$$\langle x, y \rangle \rightarrow \langle x', y', \Delta t, v, p \rangle$$

where x and y are the location where the target disappear and $\langle x', y' \rangle'$ is the location where it reappears. The value Δt is the time between the two events and p is the conditioned probability that E_2 follows E_1 after Δt, that is, $p(E_2(t+\Delta t)|E_1(t))$. This is an extension of our previous model that did not use probabilities. This probability as well as the expect location of the reappearance is updated every time a similar event occurs. To make this possible, the matching of events depends on the type of event and the location. Two events are assumed to occur at the same location if the spatial distance between the two events is below a matching radius. Note that the identity of the target is not learned. Instead, the association will be generalized to all targets.

Let $\langle \hat{x}, \hat{y} \rangle$ be the current expectation of the target position. When a target is visible, the system will use the currently viewed location as the position of the target, that is, $\langle \hat{x}, \hat{y} \rangle = \langle x, y \rangle$. However, when the target is invisible and the system expects it to appear at a particular location in the future, it will use linear interpolation between the location of E_1 and E_2 to predict the position of the target. That is, at time T after event E_1, it will predict that the target is at location

$$\hat{x} = x + \frac{T}{\Delta t}(x' - x)$$
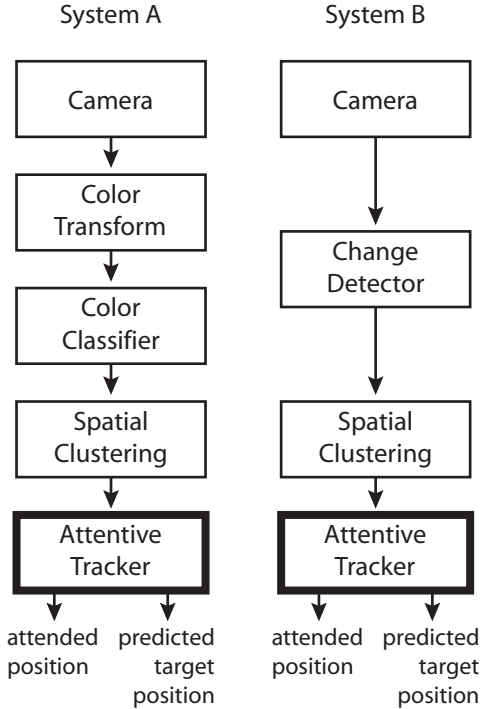
$$\hat{y} = y + \frac{T}{\Delta t}(y' - y).$$

FIGURE 1: *The two systems used to track a toy car moving on a track. A. Color based tracking. B. Motion based tracking. See text for further explanation.*

When an event leads to several conflicting predictions, the most probable prediction is selected. In addition, the certainty of the prediction will depend on the closeness to the actual event E_1 or the anticipated event E_2. This results in the calculation of certainty c as,

$$c = \alpha^{min(T, \Delta t - T)}.$$

Here, the constant α detemines how quickly the certainty falls off when the system is not sure of where the target is. The location to overtly attend is selected in the following way: If several targets are visible, the system will chose to attend to the target closet to the center of the fovea. If no target is visible, it will instead attend to the location where it expects the next target to appear.

2.2 Computer simulations

To demonstrate the ideas behind the event prediction system, we have tested it on a number of simple movies of a toy car running on a track. There is one or several tunnels into which the car disappears for a while before showing up again at the other end. A frame from the scene is shown in Fig. 2.

The model was implemented as a number of interacting modules according to Fig. 1, using the Ikaros system (Balkenius, Morén and Johansson, 2007). The input was taken from a stationary camera looking at the scene. The details of the color based tar-

FIGURE 2: *The track with the car in the tunnel. The system is waiting for the car to reappear at the end of the tunnel to the right. The previously predicted path of the car (black line), the currently predicted location (white cross) and the attended location (white circle).*

get localization have been previously described by Balkenius and Johansson (2007b). The coordinates and the certainty form the localization system were used by the attentive tracker as described above.

2.3 Results

Fig. 3 shows the main signals of the system. The certainty of the target localization changes over time and drops to zero when the car disappears. When the car appears after the tunnel again, the certainty increases to its previous level. These changes are categorized as events and the relations between them are learned. As a consequence, the certainty of the attentive tracker changes so that it anticipates that the car will reappear. This is seen in Fig. 3b where the certainty does not fall to zero the second time the car disappears. Instead it gradually decreases only to increase again before the expected reappearance of the car. The final graph shows the expected horizontal location of the car. The first time it disappears, the system does not know what will happen and the predicted location stays where the target was last seen only to make a quick jump to the other side of the tunnel when the car appears. The second time the car enters the tunnel, the prediction will instead follow a smooth path from between the two ends of the tunnel. This reflects covert attention that tracks the target by interpolation even when it is not visible. On the other had, the gaze will immediately move to the location where the car will appear.

The reason why the gaze moves immediately to the end of the tunnel is that the system does not show any interest in the other parts of the scene. The gaze had not moved immediately to the other side, had there been other potential targets in the scene, but they are made essentially invisible by the target localization part of the model.
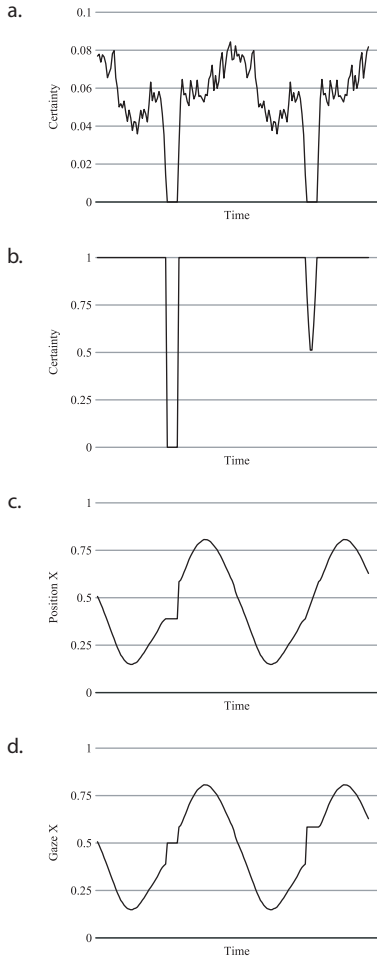
FIGURE 3: *The certainty of the car localization (a) and target prediction (b) during two laps around the track. Events are detected when the top curve drops to zero or return to a higher value. Note that the certainty of the tracker increases after the first observation of the car entering end exiting the tunnel. (c) The predicted horizontal position of the car. During the first time around the track, the system is not able to predict the location of the car when it is hidden by the tunnel, but the second time around it will fill in the expected location and the predicted location changes smoothly even when the car is not visible. (d) The attended location. The first time the car disappears, the attention will be directed to the center of the image, but the second time this happens, the gaze will move to the end of the tunnel in anticipation of the car.*

3. Anticipation in a multi-agent scenario

The presentation above only described the case with a single target. However, the system is capable of tracking several simultaneous targets. It can also be used as a part of a cooperative tracking involving several robots. These extensions are described in this section.

3.1 The architecture of a complex anticipatory system

The complex anticipatory system consists of four parts for learning to anticipate visual event together with the robot simulation Fig. 4. The system used in this paper has been used for other experiment studying anticipatory behaviours (Johansson and Balkenius, 2007).

Robot Simulator The first module is a multi-robot simulator, which includes a large number of modules both for simulation and robot experiments. The whole system have been developed primarily for real robots, but can also be run as a pure simulation. The system can be adjusted to simulate different types of robots although here we use the kinematics of the e-puck robot. The e-puck robot is a small two wheel robot. It has a maximum speed of roughly 12 cm/s and uses differential steering. We have six e-puck robots in our lab and we have chosen six agents to make it convenient to compare the result from simulations and robot experiments.

Visual Filtering The second module, visual filtering, is used to curtain the visual field for each agent. Instead of a total knowledge of the environment and agents within it, each agents is only allowed to use the information from locations where they direct their attention. The visual field is projected as a triangle in front of the agent (Fig. 5). If something appears within this visual field, the agent can register this event and use it to chose appropriate actions. In this scenario, events like agents disappearing or appearing are forwarded to the attentive tracker module. The visual filtering module also detects the obstacles surrounding the agent and whether an obstacle is blocking the direct line of sight between the agent and a target. In the case of a direct line of sight, it will forward the exact position to the rest of the system and if there is an obstacle blocking the line of sight it will not perceive the other agents positions.

The width of the visual field can be adjusted to simulate different types of cameras or eyes. Usually robots have cameras with a visual field of 50 to 80 degrees which contrasts with humans and animals where the visual fields around of 180 to 360 degrees are common. The input for the visual field module is the head direction of the robot.

In the case of the e-puck robot, there is no actual robot head. To use the on board camera in a e-puck experiment, it is necessary to direct attention using whole body rotation, and as a consequence, the navigation system and the attention system will interfere with each other. When the attention system has higher priority than the navigation system, the

robot will stop and turn in that direction until the navigation system will regain control. To overcome this and other issues using the on-board camera, like insufficient memory to receive the whole image and slow Bluetooth communication for sending images, we instead use an overhead camera to track robots and obstacles. In this paper we simulate the overhead camera part of the system to obtain the same type of input as we would have in an experiment with robots.

When the overhead camera is used instead of the on-board camera, either in simulation or in reality, we can equip our agents with virtual heads that allow the attention system and the navigation system to work more independently from each other. Still, the attention system could override navigation system if the robot have to move to be able to perceive the attention area.

The visual filtering module also provides information about events to the attentive tracker module which learns the association between agent disappear and appear. This module also has an attention output of the potential region which an agent may appear.

Attentive Tracker The attentive tracker works as described earlier, but is in addition able to handle several simultaneous targets. The first time a target disappears, the attentive tracker does not have any association between this particular position and where it may reappear. The system suggest to stay focused on the region where the robot disappeared. Every suggested attention area is associate with an certainty level of how likely the agent will appear at that point. This level decreases exponential, over time while the target is unseen. If an unassociated event is triggered the system will focus the attention where the target disappeared for a while and then lose interest and other attention regions will be activated instead.

Epistemic Actions When the attentive tracker determines with high certainty that a target will appear, the epistemic action module tries to direct the agent's attention to this region. The only epistemic action currently implemented is head movements although the system could benefit from more complex search behaviors and other actions involving change of head direction or agent movement to a certain point at a certain time etc.

If the certainty that an agent will appear is high from the attentive tracker module, the epistemic action module will instruct the virtual robot head module to turn in the direction of the interesting area. The robot head will smoothly track a visible target until it is concealed by an obstacle and stop its tracking at the point where the agent disappeared
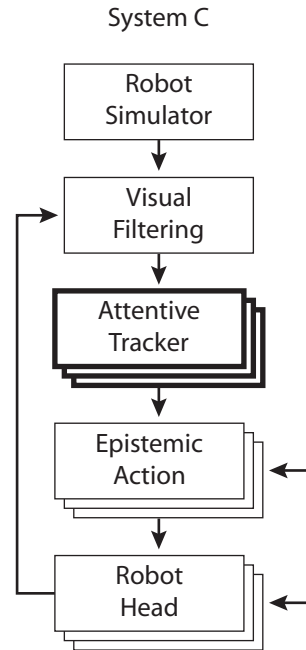
System C

FIGURE 4: *The complex anticipatory system uses four parts for learning to anticipate visual events. The first module filter the global visual input to local visual input for each robot. Attentive tracker learns the relation between a disappear event and an appear event. The last two parts handles the attention direction of the agent. The robots simulation include necessary modules for simulation of the e-puck robot is used to provide robot simulation for the four parts.*

and when the time prediction of the approach of the taget, the attention will shift to where the agent will appear.

In the case of low certainty, the epistemic action module starts to explore its environment although it will shift back the attention as soon as its time for the agent to appear.

Robot Head The last module included in the attention system is the robot head module. This module simulates the head direction of the agent and this direction is used by both visual filtering module and the epistemic action module.

3.2 Computer simulations

To test the attentive tracker in a more complex environment we simulated 6 agents, two guards and four scouts. The task for the scouts were to predict the positions of the two guards that moved along a regular periodic path through the environment. With this simulation, we tested how the different scout position, visual field widths and complexity of the world influences the prediction performance of the guards.

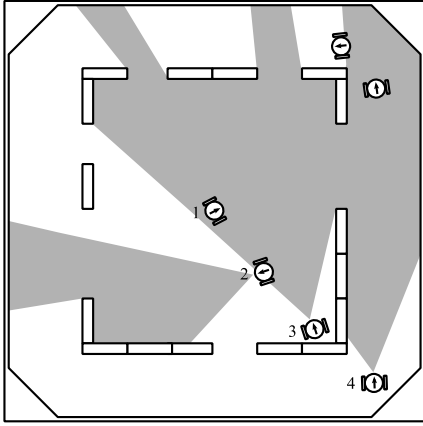All the experiments in this section uses a simulation of the e-puck robot. None of the on board sen-

FIGURE 5: *Two robots in the upper right corner are the guards whom the other robots tries to predict.*

FIGURE 6: *Performance of the prediction of the location of the guards for each lap. Each line shows the increase in performance for one individual robot.*

FIGURE 7: *The learning time until a good prediction can be made varies on the width of the visual field. A wider visual field gives faster learning to predict visual events.*

FIGURE 8: *Learning rate for different sizes of the openings of the occluders: (solid) narrow openings, (dashed) medium openings, (dotted) wide openings.*

sors are used in the experiment. Instead all sensory input comes from the simulated overhead camera.

Three different experiments were conducted using a single agent. The first experiment investigated the mean error between actual position and the predicted position for each lap around the environment for different placements of the scouts. In the second experiment we varied the width of the visual field to see how it influenced the mean prediction error. In the last experiment, the complexity of the environment was varied.

All experiments uses the Ikaros modules describe earlier and illustated in Fig. 4. The scouts are placed at different locations from the middle of the scene to the lower right corner Fig. 5. The scouts do not move away from their initial positions during the experiment, but they are able to move their heads to track any moving target.

When we varied the position of the scouts the prediction also varies (Fig. 6). The best prediction is made by scout 1 and the worst prediction is made by scout 3. All robots has learnt the prediction well only after 3 laps and after 5 laps not much improvements can be made. The direct value of the graphs are not comparable because the optimal mean error between the prediction and the actual position of the guards, is different for different placement of the agents, instead one have to look at how the graphs changes over time. If a robot has a larger area that it can not cover using vision, the prediction mean will increase.

When changing the width of the visual field for one of the agents the result indicated that a wider visual field will produce smaller mean prediction error in comparison with a narrower visual field (Fig. 7). With 360 degrees visual field, only one lap is necessarily to get an optimal prediction of the guards. Suprprisingly, even with a narrow visual field, the learning time until a fairly good prediction is obtained is short. After 3 laps, the performance be-
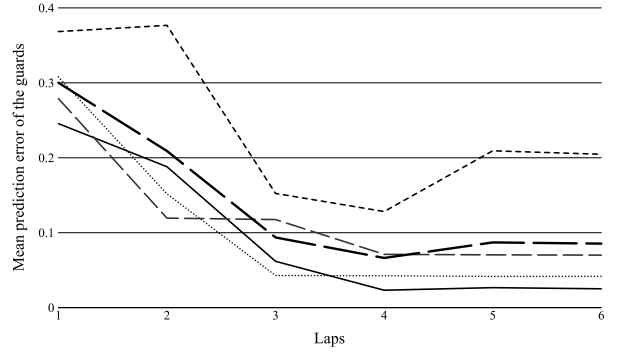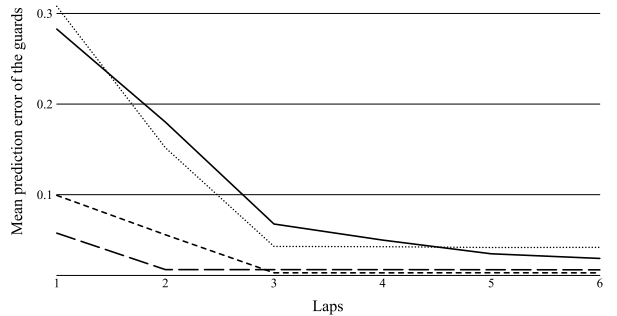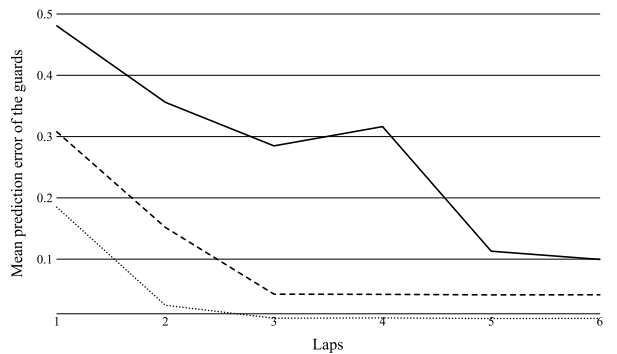
tween the different visual fields are vary close to each other.

That increasing the openings between the obstacles improves the prediction was shown in the last experiment (Fig. 8). Also the learning time decreases when larger openings were used.
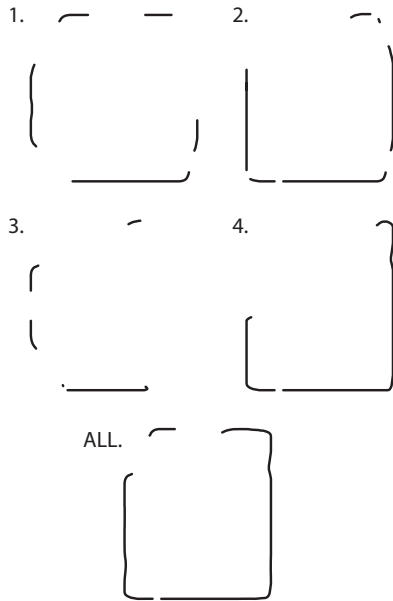
FIGURE 9: *The top four plots show the predictions of the individual scouts of a single guard. The lower plot is the combined prediction of all the scouts.*

Cooperating agents   Finally we let the scouts co-operate to see how the prediction would improve. The measurements for this experiment are the mean prediction error and also the certainty of a target position. The same set-up as in previous experiments are used with the difference that in this experiment the robots different prediction is weighed together depending on their certainty of the target position. Fig. 5 show how the cooperative visual field.

Fig. 9 shows how the cooperative prediction increases comparing to a single robot prediction. The result from this experiment indicates that the scouts will benefit if they cooperate and share their knowledge of the world. In this experiments no feedback on how well the prediction worked was used. This feedback is necessarily for an optimal merging of the predicted target locations.

4.   Discussion

We have presented the architecture of an event prediction system that can be used to control overt attention as well as covert tracking of multiple targets. The current implementation has focused on event prediction in realistic situations. Our first example was a tracking task with video input depicting a scene with a moving car and the system was shown to quickly learn to predict where the car would reappear after it disappeared into a tunnel. The second example was a task with multiple robots that must observe the behavior of other robots. Again, the proposed system is quickly able to learn to track multiple robots.

Unlike our previous models (Balkenius and Johansson, 2007a), we did not include systems for short term prediction of motion. Instead we focused on the role of event associations and the difference between the predicted location of the target and the overtly attended location in the scene. We showed that by decoupling covert prediction of target location from overt attention to a target, it becomes possible to track several targets in an efficient way even when they may be partially occluded behind obstacles.

By making the visual field larger, either by a better placement in the environment or by having an initially wider visual field, the learning time needed to make a good prediction decreases although even with a smaller visual field a fairly good result is obtained after only a short time. However, to get an optimal performance, a larger visual field should be used. Other ways to improve the observations in the environment is to use more sophisticated attention mechanisms. It would be possible to use an attention system that will remember certain places where targets usually are seen. This would make it possible to prioritize the search to avoid looking at places where the agent can not be seen eg. in walls. The attention system could also instruct the agent to move to a better position to increase the size of the visual field.

We have also shown how a group of robots with limited attentional resources, but different fields of view, can cooperate to track targets over larger areas. To be able to benefit most in the cooperative scenario, a certainty value that indicate the correctness of the prediction is necessary. Without this value, agents with incorrect predictions will lower the overall correctness.

Although the use of several cameras that collectively track multiple targets has been thoroughly studies in computer vision (Stauffer and Grimson, 2000, Ercan et al, 2007), our goal here is to investigate how our model of infant attention can be used in such as situation. The problem of tracking targets through occlusion has received much interest in computer vision. For example, Stauffer (2005) used Adaptive Background Mixture Models to detect objects and learned Transition Correspondence Models to track object between cameras. Pan et al (2008), propoed a method that is able to track object through short as well as long term occlusion. Other systems make use of various forms of reasoning to disambiguate the scene after occlusion (Bennett et al, 2008, Mottaghi and Vaughan, 2007).

The current system generalizes the behavior of a target behind an obstacles to all other target objects. Although this is appropriate for the current scenarios, it is clear that it is not always correct. In the future, we would like to extend the learning system to make it possible to generalize in different ways in

different cases. For example, a behavior particular to a single target object should not be generalized to all other objects. We envision that some form of context dependent learning could be useful in this situation (cf. Balkenius and Winberg, 2008). The learning algorithm for the predictions share many assumptions with the DRAMA architecture (Billard and Hayes, 1999). Learning relies on event detection and associations code both a time-delay and a confidence in the form of a conditioned probability.

The suggested system is able to predict the future location of multiple targets. It has recently been suggest that this ability is not present in humans (Keane and Pylyshyn, 2006). However, this contrasts sharply with the evidence that shows that even infants are able to predict the future location of a single attended target (Rosander and von Hofsten, 2004). It is possible that different mechanism are used for a single actively attended target and several passively tracked target. Another possibility that is suggested by the present system and that seems to be consistent with the experimental data is that prediction is used only for interpolation between two known events, but not for extrapolation of covertly attended targets.

Acknowledgements

References

Balkenius, C. and Johansson, B. (2007a). Anticipatory Models in Gaze Control: A Developmental Model. Cognitive Processing, 8, 167-174.

Balkenius, C. and Johansson, B. (2007b). Finding Colored Objects in a Scene. LUCS Minor 12.

Balkenius, C., and Winberg, S. (2008). Fast Learning in an Actor-Critic Architecture with Reward and Punishment, In Holst, A., Kreuger, P., and Funk, P. (2008). Tenth Scandinavian Conference on Artificial Intelligence: SCAI 2008 (pp. 20-27). Amsterdam: IOS Press.

Balkenius, C., Morén, J. and Johansson, B. (2007). System-level cognitive modeling with Ikaros. Lund University Cognitive Studies, 133.

Billard, A. and Hayes, G. (1999). DRAMA, a connectionist architecture for control and learning in autonomous robots. Adaptive Behavior Journal, 7, 1, 35-64

Broadbent, D. E. (1958). Perception and communication, Pergamon Press, London.

Cavanagh, P. and Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. Trends in Cognitive Sciences, 9, 7, 349-354.

von Hofsten, C. and Rosander, K. (1997). Development of smooth pursuit tracking in young infants. Vision Research, 37, 13, 1799-1810.

Johansson, B. and Balkenius, C. (2007). An Experimental Study of Anticipation in Simple Robot Navigation. In Butz, M. et al. (Ed.) Anticipatory Behavior in Adaptive Learning Systems: From Brains to Individual and Social Behavior. LNAI, 4520, Springer-Verlag.

Keane, B. P. and Pylyshyn, Z. W. (2006). Is motion extrapolation employed in multiple object tracking? Tracking as a low-level, non-predictive function. Cognitive Psychology, 52, 346-368.

Oksama, L. and Hyoumlnauml, J. (2004). Is multiple object tracking carried out automatically by an early vision mechanism independent of higher-order cognition? An individual difference approach. Visual Cognition, 11, 5, 631-671.

Piaget, J. (1954). The construction of reality in the child. London, UK: Routledge & Kegan-Paul (Originally published in French in 1937).

Prem, E., Hörtnagl, E. and Dorffner, G. (2002). Growing Event Memories for Autonomous Robots. In Proceedings of the Workshop On Growing Artifacts That Live, Seventh Int. Conf. on Simulation of Adaptive Behavior, Edinburgh, Scotland.

Pylyshyn, Z. W. and Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. Spatial Vision, 3, 3, 1-19.

Rosander, K. and von Hofsten, C. (2004). Infants' emerging ability to represent occluded object motion. Cognition, 91, 1, 1-22.

Stauffer, C. (2005). Learning to Track Objects Through Unobserved Regions. IEEE Workshop on Motion and Video Computing (WACV/MOTION'05), 2, 96-102.

Stauffer, C. and Grimson, W. E. L. (2000). Learning patterns of activity using real-time tracking. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22, 8, 747-757.

Wentworth, N. and Haith, M. M. (1998). Infants' acquisition of spatiotemporal expectations. Developmental Psychology, 34, 2, 247-257.

Yantis, S. (1992). Multielement visual tracking: Attention and perceptual organization. Cognitive Psychology, 24, 3, 295-340.