Balkenius, C., Kopp, L. (1996). Visual tracking and target selection for mobile robots.
In Jörg, K.-W. (Ed.) *Proceedings of EUROBOT '96* . IEEE Press.

Visual Tracking and Target Selection for Mobile Robots

Christian Balkenius
Lars Kopp

Lund University Cognitive Science
Kungshuset, Lundagård
222 22 Lund, Sweden
christian.balkenius@fil.lu.se
lars.kopp@fil.lu.se

Abstract

This paper describes how tracking and target selection are used in two behavior systems of the XT-1 vision architecture for mobile robots. The first system is concerned with active tracking of moving targets and the second is used for visually controlled spatial navigation. We overview the XT-1 architecture and describe the role of expectation-based template matching for both target tracking and navigation.

The subsystems for low-level processing, attentional processing, single feature processing, spatial relations, and place/object-recognition are described and we present a number of behaviors that can make use of the different visual processing stages.

The architecture, which is inspired by biology, has been successfully implemented in a number of robots which are also briefly described.

1. Introduction

Many uses of active vision can be divided into the two tasks of target selection and tracking. This is especially true for visually guided goal-directed behavior. This article describes the two most important goal-directed visual behavior systems in the XT-1 (eXpectation based Template matching) architecture [5]. This architecture is an attempt to design a biologically inspired model of a number of visual tasks such as orienting and anticipatory saccades, smooth pursuit, landmark and place recognition, and visually guided locomotion.

The first, of the systems described below, implements object tracking and is currently used in the LUCS Active Stereo Vision Head to find and track moving objects. The second behavior system uses visual tracking to carry out stimulus-approach (S-A) behavior [2, 7]. This second behavior-system is currently used in the POLUCS robot as an important component in spatial navigation. We believe that these two behavior-systems, which share many structural features, can be used to great utility in a number of mobile robot tasks. Below, we describe how target selection and tracking is implemented in each system.

2. Overview of the XT-1 architecture

The architecture can be divided into five conceptual levels: low-level processing, attentional processing, single feature processing, spatial relations, and place/object-recognition (See figure 1, and [5]). At each higher level the representations become more complex, but the processing is fundamentally heterarchical: the information flow is both bottom-up and top-down, as well as lateral.

The first level is concerned with low-level preprocessing of video-images. A scale-space pyramidal edge-detection constitutes the first stage at this level. In the second stage, the difference between successive edge images is used as a quick-and-dirty motion detection.

The second process level deals with attentional processing based on the input from the first level. A primitive attention module directs the attention of the tracking subsystem to sudden motion in the environment and triggers an orienting saccade toward it. When the navigational subsystem is disengaged, this primitive attention system is used to select targets for the tracking system. This module is inhibited while the camera-head is moving. A second parallel system directs attention to regions with potentially good features in the image. At the higher levels, these regions are used as candidate landmarks.

Unlike the two previous levels which perform global computations, only local features are processed at the third level. The single feature processing is applied to regions of the image that have been selected, either by the attentional systems, or by top-down influences from higher levels. The feature-correlator is the central component of this level and is used both to compute optic flow and to locate landmark and target features in the image. A search-field module is used to control where in the image it is fruitful to compute local feature correlation. The role of this module is to reduce the amount of computation required by the system.

At the fourth level, the spatial relations between individual features are used to represent landmarks in the navigational subsystem. Such collections of features can also be sent to the tracking subsystem when the robot

needs to pursue a goal. When the tracking system acts on its own, the optic flow calculated at the lower level controls a segmentation process where a region of homogeneous motion is selected as target.

Finally, at the fifth level, the angular relations between landmarks come together to form the representation of places. Such relations can be seen as second order-spatial relation, i. e., relations between collections of features which themselves are grouped with their spatial relations. Note that using this scheme, no object recognition or complicated segmentation is necessary to categorize a place. To a first approximation, it appears that also object recognition is a process at this level.

The top-down influences between the levels of the architecture are used to direct the computations at each lower level toward regions of the image with the largest expected amount of information. For example, the place-module generates expected landmarks to look for in the scene. The landmark module, in turn, generates expectations of the features that are likely to found in the image. This speeds up the computations considerably in most cases.

3. Object tracking

Two processes are involved in tracking a moving target. The first is to find the target in the image, and the second is to control the motion of the camera head. Many different approaches are possible in both these areas. In our earlier systems, the target was found and tracked using simple motion detection [3, 4]. This requires that the camera must be still when the target is to be detected since it cannot distinguish between target motion and self motion. The requirement that the camera must be still forced us to use a design where the head would follow the object using a succession of small saccade motions. In between the saccades, the camera was still and attempted to spot the target again. The length of each saccade was proportional to the distance from the centre of the camera image to the centre of the target. Given that the delay between each sampled image is not to long, this system would successfully track a moving object.

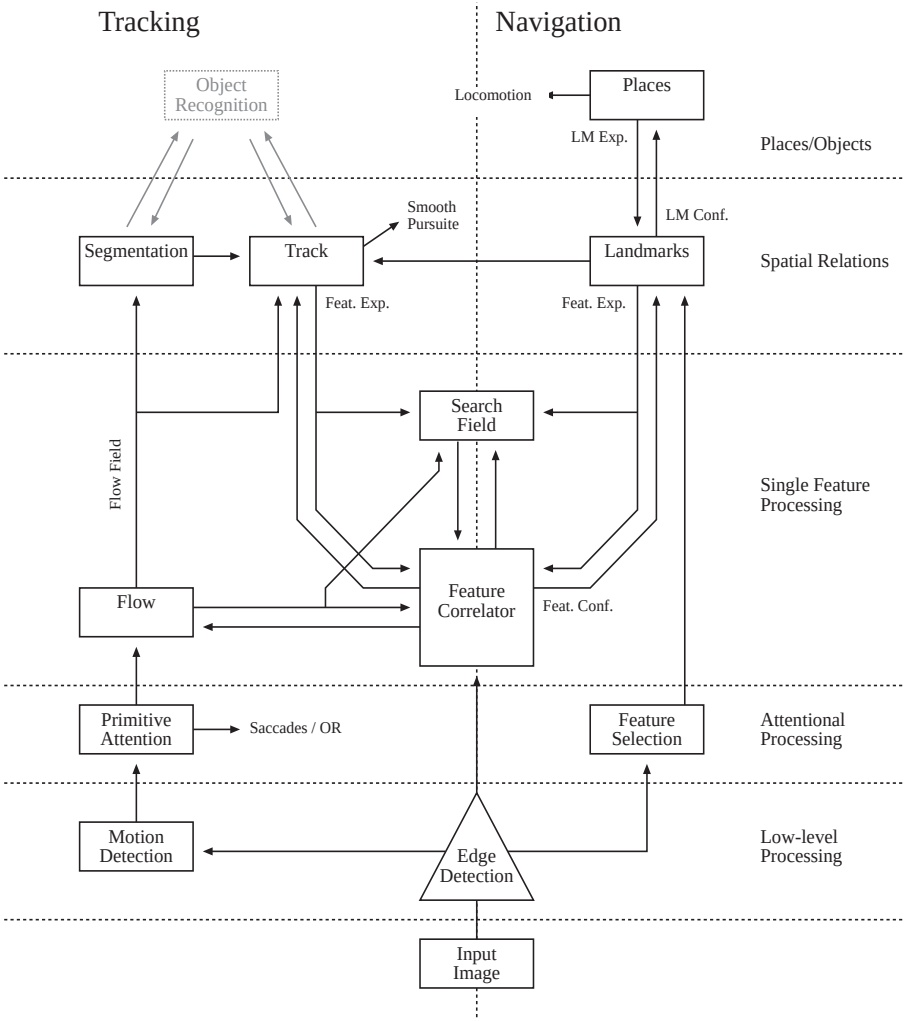

Figure 1. Overview of the XT-1 vision architecture.

If the target is moving sufficiently fast compared to the speed of the camera, it is possible to track a target to some extent despite the self motion. This approach was used in our earlier LUCSOR robot to make it approach a waving hand or to pursue a moving person [3]. When many moving objects are present, however, this system more or less breaks down. It also generates very jerky movements in the camera head since the head is constantly stopped only to be started again a moment later.

A much better tracking obviously results if the camera follows the target continuously without stopping. In this case, however, it will no longer be possible to use motion detection to find the target. This is the approach taken in the XT-1 architecture where a number of interwoven processes are required to find and track a moving target.

3.1. Target selection

Studies of biological systems suggest a number of ways to select targets for visual tracking. The most important selection behavior is based on the so called orientation reaction [2, 6]. This is a reaction found in most animals with some level of sensory sophistication. Basically, the role of this reaction is to direct the sensory apparatus of the animal toward the source of an unexpected event in the environment. The simplest form of visual orienting reaction is triggered by the on-set of motion in the visual field. In animals with movable eyes, the reaction can be divided into two components [8]. First, the eyes make a fast saccade toward the center of motion in the image. Subsequently, a recognition process that can be more or less complex decides whether the source of the motion information is to be considered interesting or not. If not, the eyes return to their initial position. Otherwise they remain at the potential target.

Second, a decision is made on whether the head should turn toward the motion or not. If the source of the motion is too far away from the forward direction of the head. The head is turned while the eyes compensate for the head movement until the target is straight ahead of both the eyes and the head.

In some cases, this sequence of movements is followed by a more or less complete body turn toward the direction of the head. This idea was used in the LUCSOR robot to make it follow a moving person [3]. First the two-degrees of freedom head would track the visual motion, and thereafter, the robot body would track the direction of the camera head during forward locomotion.

The result of the orienting reaction is thus to move a potential target to the center of the visual field. Typically, this reaction is multi-modal [8]. For example, an unexpected sound may trigger an orienting reaction that will direct the eyes toward the source of the sound. Similarly, in an animal with movable ears, a sudden visual stimulus may turn the ears in its direction. [1] describes a neural network model of how the required multi-sensory mappings can be learned automatically.

In the object tracking part of the XT-1 architecture, target selection is performed in a way similar to its biological counterpart. After the initial edge detection stage, a primitive motion detection is performed by comparing two successive edge-images (See figure 1). The amount of change at each pixel is used to construct a primitive attention field across the image. The amount of attention at a region of the image increases with increased primitive motion. In the subsequent processing stages, the amount of computational resources allocated to each region of the image is proportional to the amount of primitive attention it receives.

The primitive attention map is also used to generate orientation movements toward a potential target object. The selection of the target for the orienting movement in a multi-sensory robot is further described in [1].

Once the target is in the center of the visual field, it is possible to segment the moving object from the background. In the XT-1 architecture, this is done using an attention-driven visual-flow mechanism.

First, the direction of the visual flow is calculated at the locations in the image where a sufficient level of attention has been allocated. The optic-flow computations are thus data-driven by inputs from a low-level primitive attention system. The optic-flow computations are based on a correlation method where features are correlated in a restricted search area, the search field, between two successive images. The search field is intimately connected with expectations since expectations of the target location and movement govern the shape of the search field. For example, when an object moves fast in a certain direction, the search field enlarges in that direction. This is an adaptive regulation which makes it possible for the tracking process to follow fast moving objects.

Second, in the segmentation module, local motion vectors are grouped together based on proximity and direction to form motion segments. In this stage, optic-flow is used to make figure-from-ground separation. This process integrates optic-flow information and categorize regions of the image with coherent motion. To do this, a neural network classifies motion-directions into eight categories. Neighbouring regions with the same direction preference are evaluated as a group and the largest group is selected as the object of interest.

In future versions of the architecture, we intend to include a subsystem above this level which will categorize the selected segment as a particular object, but much further research is necessary before this is possible in more than a few restricted cases.

3.2. Tracking

When a target has been selected, the next task is to track it with the camera head. In the current implementation of the architecture, only a single feature is used to represent the object (but see section 5 below). We have experimented with object representations using collections of features, but since those attempts did not increase the performance of the tracking system, we have temporarily stayed with the single feature representation.

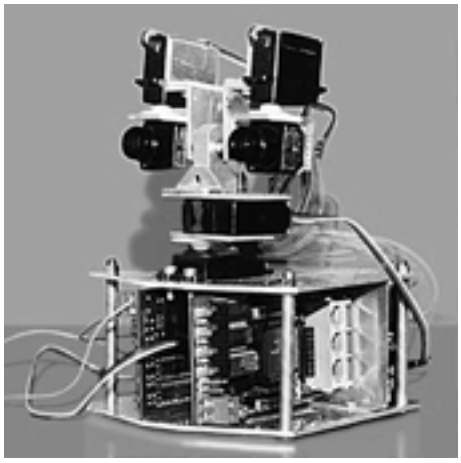

Figure 2. The LUCS Active Stereo Vision Head. The head has four degrees-of-freedom (pan, tilt, 2¥vergence) and is controlled by the vision-architecture described in the text.

The tracking in the input image is a simple process. For each new input image, the previous object feature is searched for in the image using a correlation method. The previous location of the target controls the size and location of the search field used by the feature correlator. This information constitutes a top-down influence on the correlation process.

The complicated part of the visual tracking lies in the control of the camera head. The LUCS Active Stereo Vision Head is a four degrees-of-freedom camera head for stereo vision (figure 2). The present head is the third camera head developed at Lund University Cognitive Science, and the first to use stereo vision. The head is very cheap and is intended for use with a low cost computer system such as a 486 with fairly limited capabilities (The total cost of the prototype is less than 6000 SEK ≈ $900). This has constrained the possible designs of both the hardware and software and it should be kept in mind below that the computer system used is very modest.

Currently, only two of the four degrees-of-freedom are used for object tracking. To simplify the tracking process, only the pan and tilt servos are used although the control system will be extended to the vergence servos in the near future. The positions of the servos are updated at a frequency of 25Hz. In theory, it would be possible to control the positions of the servos directly, but since the mechanical construction is fairly unreliable, feedback control is used instead.

It turned out to be impossible to use a single control rule both when the target is moving slowly and when it is jumping over the image at a high speed. Again, we used the biological counterpart as an inspiration. Studies of the human visual system have shown that object tracking is often divided into two types of eye movements: smooth pursuit and saccades. In smooth pursuit, the eyes track the target with almost constant velocity. Using saccades, however, the eyes jump in small steps from the previous location of the target to the next.

The selection of the appropriate behavior is determined by the distance to the target. If it is close to the middle of the retina, smooth pursuit is used. When it moves too far off the center of the image, a saccade is generated which tries to make up for the lag in the tracking. This suggested to us the use of a switching control strategy where smooth tracking is performed as long as the target is close to the middle of the image. When it moves out of this region, a saccade is performed that tries to directly place the target in the center of view.

To summarize, the target selection and tracking divides the visual field into three regions: peripheral, intermediate, and focal (figure 3). In the peripheral region, an orienting reaction is generated towards the center of motion in the image as described above. The other two regions are used in object tracking to select whether to perform a saccade or smooth pursuit. When the target is in the focal region, smooth pursuit is performed, and if it moves to the intermediate region, a saccade is generated.

4. Approach and pursuit behavior

Object tracking in itself is of no other use than to keep the target close to the center of the video-image. Tracking becomes more interesting, however, when it is used to control the locomotion of a robot. Given that a robot head which is mounted on a mobile robot tracks an object, the direction of the camera head can easily be used to guide an approach behavior toward the object. Behaviors of this type are called stimulus-approach behaviors since they make the robot, or an animal, approach a specific stimulus [2, 7]. They are the simplest examples of goal-directed behaviors.

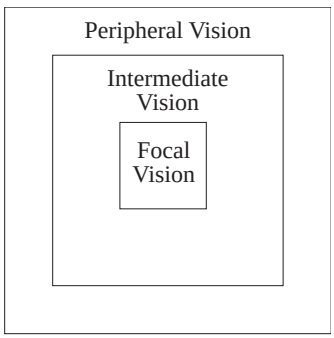

Figure 3. The three regions of the input image. Peripheral vision. Sudden motion in this region will generate an orienting reaction toward the center of motion. Intermediate vision. A target found in this region will generate a saccade toward its location in the image during tracking. Focal vision. A target in this region can be followed using smooth pursuit.

A special kind of stimulus-approach behavior results when the whole mobile robot moves toward a moving target. Such a behavior can be called a pursuit movement. In the LUCSOR robot, pursuit behavior was implemented using a distributed scheme [3]. The camera head would track the target independently of the body of the robot. The body, in turn, would receive information of the current direction of the head relative to the body and would try to track that direction.

A similar method will be used in the stereo head in the future to control the movements of the individual servos. First, the target is tracked in the image from the dominant camera. Second, the vergence servo for that camera tries to track the location of the target. Simultaneously, the second camera will track the feature in the center of the image from the dominant camera. This will take care of the vergence control. Third, the pan servo will track the average direction of the two cameras. Finally, the body of the mobile robot will track the direction of the camera head.

Using this scheme, each servo can be controlled in a distributed manner with only very limited knowledge about the other parts of the system. To avoid oscillations, each subsequent stage must move at a decreasing speed. The body must move slower than the head, which in turn must move slower than the cameras.

In an environment with sufficiently many potential targets, stimulus-approach behavior can be used for navigation since any locomotion sequence can be seen as a sequence of stimulus-approach behaviors [2, 7]. This idea is used in the navigational system of the XT-1 architecture. This requires that the target selected at each time is one that makes the robot approach its final goal.

5. Landmark selection and tracking in spatial navigation

In the XT-1 architecture, the most complex tracking is used within the navigational system. Here, the targets used are landmarks rather than objects. We define a landmark as any collection of visual features with constant, or almost constant, spatial relations to each other and the environment. A collection of features together with their spatial relations makes up an elastic template for each landmark.

Like object tracking, spatial navigation rests on the two processes of target selection and tracking. This part of the architecture is currently running in the POLUCS robot (figure 4) which can successfully move toward an initially invisible goal using only visual input. Like the camera head described above, very limited computational resources are needed.

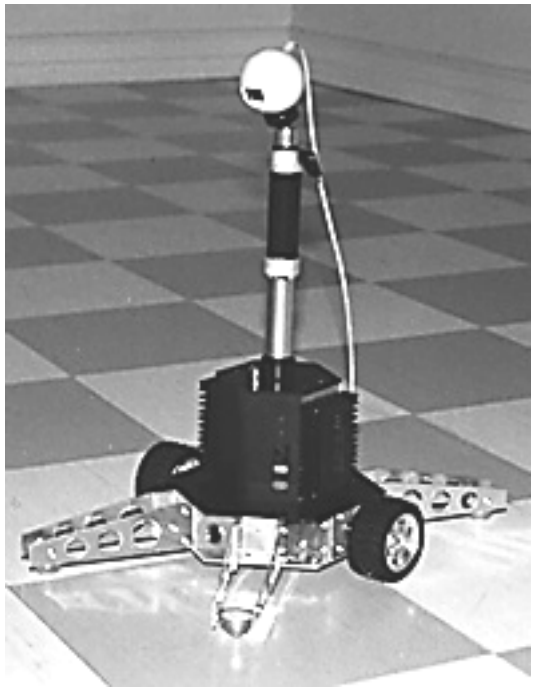

Figure 4 The POLUCS robot that navigates using visual landmark recognition and sequences of stimulus-approach behaviors.

5.1. Landmark selection

The target selection in the navigational system is more complex than that for the object tracking system. While the previous system only uses a single feature as target, the navigational system uses several landmarks, that is, collections of features, at each time. These landmarks are selected in one of three ways.

First, the robot always keeps one current target landmark that has already been selected. This is the target toward which the stimulus-approach behavior directs the robot. If this target landmark is lost for some reason, the robot performs an orienting behavior which tries to locate any known landmark around the robot. When and if such a landmark is found, it is selected as the current target.

Second, at each location, the robot expects to find other landmarks closer to the goal that it can approach. A set of such expected landmarks are always generated by the navigational system and if any of these potential targets are found in the image, it is selected as the new target.

Third, the expected landmarks are stored together with their expected angle from the current forward direction of the robot. When an expected landmark is too far to the side to be found within the input image, an anticipatory saccade is performed toward the expected angle of the landmark. If it is found, it is selected as the new target. This behavior makes it possible for the robot to turn 90 degrees or more when necessary.

5.2. Tracking landmarks

The landmark tracking in the navigational system differs in many respects from that in the object tracking system. Most importantly, a landmark is made up of many features together with their spatial relations. As a consequence, the feature correlator has a much more resource demanding task to perform. To calculate the location of the landmark in the video image, it is necessary to look for a whole number of features. Again, expectations of the feature locations are used to speed up the process. When the first feature is found in the image, its location constrains the search for the next feature, and so on. Using this scheme together with the target selection mechanisms described above, we can comfortably run the landmark tracking system in real time on a standard Pentium computer.

In the current implementation of the system, the camera on the mobile robot is stationary. Instead, the whole body of the robot turns to track a landmark. This works well when the target landmark is fairly straight ahead, but is very cumbersome when anticipatory saccades need to be performed. In future systems, the robot will be equipped with a movable head similar to the one described above which will allow the robot to look around and navigate much faster.

By connecting several landmarks in series, the robot is able to perform a sequence of stimulus-approach behaviors that leads it from any initially learned location to a specified goal. Currently, the system has only been tested with a single track from start to goal, but the extension of the system to more advanced spatial representation will not increase the burden of the visual system very much. Balkenius [2] describes the neural network model of spatial navigation that will be incorporated in the POLUCS robot in the future.

6. Discussion

Both the object tracking system and the navigational system shares many important properties. They can both be divided into subsystems for target selection and tracking, but these systems are implemented in rather different ways. Nevertheless, the similarity between these two systems suggest that it may be possible to use the same mechanisms for both processes.

For example, the multi-feature aspect of the landmark tracking system could probably be adapted for object tracking. While this may seem obvious, we have not yet managed to use a multi-feature approach to object tracking. The reason for this is that the environment around a moving object does not move with the object.

Since there is no way for the tracking algorithm to know which part of the features belong to the target and which parts belong to the background, they can easily stick to the background instead of moving with the target. This makes it very likely that the features fall off the target when it turns or passes over, for example, a textured background. To solve this problem a more advanced segmentation is required. This is not a problem in the navigational system since a landmark has the property that it does stick to the background. If not, it will not function very well as a landmark. For example, it is not a very good idea to use a person standing in a room as a landmark.

On the other hand, it is not possible to use motion information to detect landmarks since the navigating robot is constantly in motion. To use the same mechanism for target selection in both the object tracking and the landmark tracking system, it would be necessary to subtract self-motion from externally generated motion in image. This is a fairly complex task for the very limited computational resources we are using. It may not be too complicated in theory, but in practice it is very hard. If this problem can be successfully solved, however, it would be possible to merge the two systems completely.

A different extension of the architecture would be to investigate tracking in behaviors that are not goal-directed in the direct sense described above. This includes behaviors such as corridor and wall following. In these behavoirs, the robot is guided by the tracked target but does not move directly toward it.

Acknowledgments

The support from the Swedish National Board for Industrial and Technical Development (NUTEK) is gratefully acknowledged.

References

[1] Balkenius, C., (1995), “Multi-modal sensing for robot control”. In L. F. Niklasson and M. B. Bodén (eds.) Current trends in connectionism, 203-216, Hillsdale, NJ: Lawrence Erlbaum.

[2] Balkenius, C., (1995), “Natural intelligence in artifical creatures”, Lund University Cognitive Studies 37.

[3] Balkenius, C. & Kopp, L., (1996), “LUCSOR Robot System Overview”, in preparation.

[4] Balkenius, C. & Kopp, L., (1996), “A simple object tracking algorithm”, LUCS Minor, 4.

[5] Balkenius, C. & Kopp, L., (1996), “The XT-1 Vision Architecture”. In P. Linde and G. Sparr (eds.) Symposium on image analysis, Lund 1996:

[6] Gray, J. A., (1975), Elements of a two-process theory of learning, London: Academic Press.

[7] Schmajuk, N. A. & Thieme, A. D., (1992), “Purposive behavior and cognitive mapping: a neural network model”, Biological Cybernetics, 67, 165-174.

[8] Stein, B. E. & Meredith, M. A., (1993), The merging of the senses, Cambridge, MA: MIT Press.