A framework for attention and spatio-temporal inference for an embodied visual agent

Chris Needham, Derek Magee
School of Computing
The University of Leeds
Leeds, LS2 9JT, UK

Sajit Rao
DIST
University of Genoa
Genoa, Italy

June 14, 2004

Abstract

A visual agent's ability to *generalise* a spatio-temporal protocol (a temporal sequence of spatially dependent actions) through use of spatial relationships is much more powerful than one which grounds to absolute positions. This is a report on work-in-progress. A framework for the integration of a spatial inference service with an active vision system using a steerable webcam, with three levels of attention is presented. The three types of attention combine to provide a mechanism for steering the active vision through *task driven attention*.

1 Introduction

Learning spatio-temporal protocols for a cognitive agent is our aim. Many components are needed in a system which autonomously learns from observation of positive examples of interactions between humans and objects. In many tasks, spatial relationships between objects are of great importance, and reasoning about positions of objects relative to other objects is necessary. A framework bringing together computer vision and robotics with attention and a spatial inference service is presented. Figure 1 (a) illustrates the interaction between each of these components. Within the 'attention' module, three levels of attention exist; data driven visual attention (A1), statistically learned spatial attention (A2), and symbolically learned spatio-temporal attention (A3). The 'robot and computer vision' module is linked to 'attention' through data driven visual attention (A1) one way, and statistically learned spatial attention (A2) the other. Both A2 and A3 are expectation driven models which are learned from prior information, with A3 being driven by the 'inference service'. The 'inference service' takes representations of the world from the 'robot and computer vision' and can ask it to perform a task (with a robot this could be to move something; in this task it may be to look at a particular thing and make a noise to indicate what it's looking at is the object to move). Figure 1 shows an experimental setup in our chosen test domain for the framework: building patterns of coloured blocks.

1.1 Motivation: how does a child learn?

We want to learn from observation of a task in a human-like way (i.e. like how a child may learn by observing its mother performing a task). Consider the task of arranging a

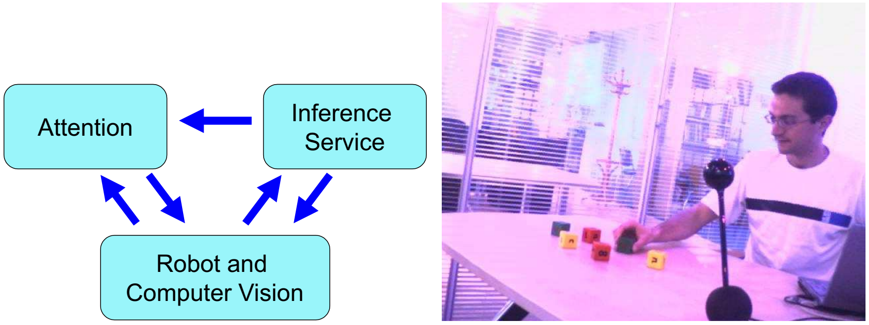


Figure 1: (a) The interaction of a spatial inference service (learning, representation, and reasoning) with a robot and computer vision system and three levels of attention. (b) Experimental setup.

set of coloured blocks in a particular pattern. It may be that the rule is to place them all in a straight line, or in two different coloured lines next to each other, or to place them clockwise in an alternating colour order around the outside of a 3 × 3 grid. A child is able to learn to to imitate its mother performing this task once it has been shown a number of examples of ‘building patterns of coloured blocks’, and can recreate these examples, and indeed other patterns which fit the same general rule, or protocol. Most importantly, they are able to *generalise* from different examples that follow a pattern. It must be noted that their are many stages which a child goes through in order to perform such a task, all of which should be considered when designing a cognitive agent:

- In the early stages, a child faces many challenges: what it can see, how to balance itself, how to understand what it can see, etc. We will not address these issues, and will assume that our system is capable of identifying objects and motion, and in addition has a mapping between its camera position and location of objects in the image.

- A child does not have a holistic view of the world. They will not remember what and where everything that they have seen is. At any given time, their vision system may not be able to see the whole field of view in which a task is performed and will focus a on particular area. This leads us to use an *active vision* system in which a steerable pan-tilt camera, which has only a limited range of view of the scene in which the task takes place.

- The focus of a child’s *attention* is largely governed by things that are brightly coloured or things that move. A visual agent needs to be able to identify salient objects in its field of view. Our visual agent identifies groups of salient features in an area of its current image and forms a representation of an object on which these features lie.

- Once attention is focused on a particular (salient) object, a number of things may happen:

  - the object moves, and the child’s eyes move to follow it (the object moves in the image, no camera movement);

- the object moves, and the child's head moves to keep it in the same place (camera moves to keep the object at the same image location);
- attention moves from the object to another salient object in view;
- a higher level non-visual process intervenes.

- A small number of (up to seven) visual markers may be used to 'remember' salient objects [7, 8]. Humans are able to store information for about a minute in short-term memory (also called working memory) and its content capacity is limited to only about seven items. A (virtual) visual marker is placed on each salient object in the scene. In this scenario, it is each object that has moved since they were all in a pile. Each object with a visual marker is tracked by the visual agent. Camera pan-tilt angles for each object are stored.

The points raised above are primarily focused on *visual attention*. In this investigation, we are looking at the interaction of different types of attention, with a hope of furthering our understanding of our own cognitive attention processes.

The way that humans behave and learn is a more complex topic; experiments are harder to devise and draw clear conclusions from. The way in which a child clusters objects into classes (red blocks, green blocks), remembers where objects may appear, forms patterns of blocks, formulates spatial descriptions between blocks, and pieces all this together to build blocks in a particular pattern are hard to specify, yet are all tasks that need to be performed by the visual agent.

2 Robot and Computer Vision

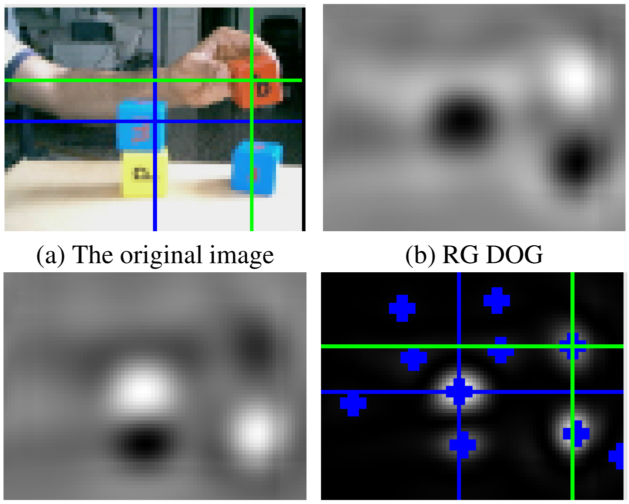
In this section we briefly explain some of the components of the active vision system which is used for the spatial template learning and can be also used for a broad range of other tasks. This allows us to observe the world and interpret what is in front of the camera. Currently no robotic hands are used for moving objects, such as those used in previous work [1], although this is an extension we wish to make. A steerable camera (Logitech QuickCam Orbit) is used, which is controlled by this system.

The Active vision processes can broadly be divided into four components:

1. **Bottom-up colour and motion saliency:** measures applied everywhere in the image.
2. **Figure-ground operations:** incorporating some top-down biases that are applied at only at the focus of attention.
3. **Active tracking behaviour:** Uses Figure-Ground operations.
4. **Marker tracking behaviour:** Maintains a short-term memory of interesting locations by using a combination of proprioceptive transformation and figure-ground operations.

2.1 Colour and Motion Saliency

Colour saliency is computed by applying a difference of Gaussians (DOG) filter on the Red-Green and Blue-Yellow channels, and taking the sum of the squares of the results. The local-maxima of the result, indicates the centres of interesting blobs. Figure 2 ht!

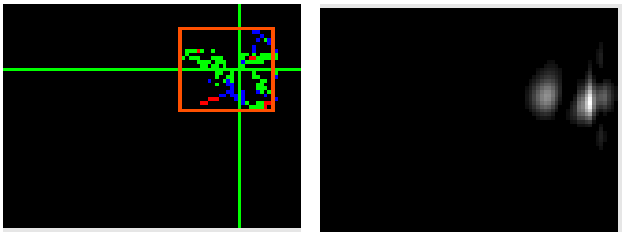


(a) The original image (b) RG DOG

(c) BY DOG (d) Sum Squared

Figure 2: Colour Saliency

ht!



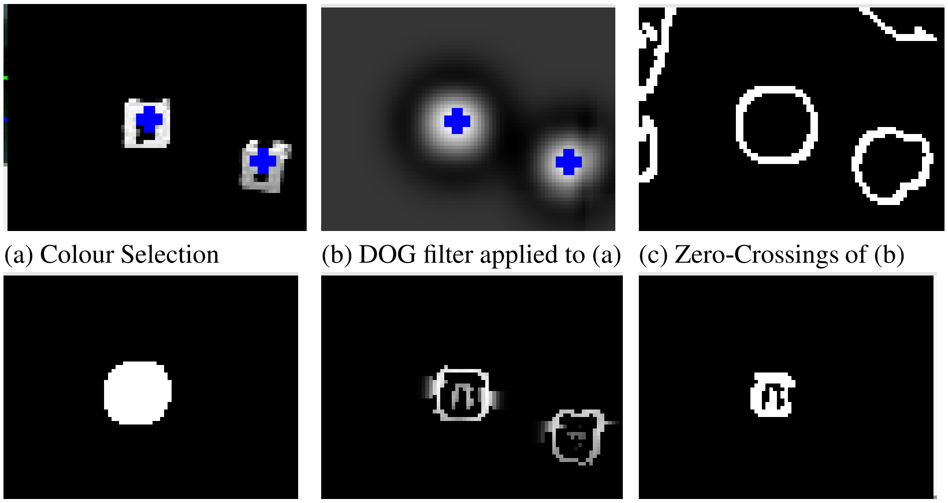
(a) Correlation based motion (b) Coloured blob motion

Figure 3: Motion Saliency

shows an example of the original image, the RG and BY DOG responses, and the sum squared image. The little blue-crosses mark local-maxima in the sum-squared image, i.e. centres of salient blobs.

Motion saliency is computed by two measures: One is a simple correlation-based optical flow, followed by a connected-components and bounding box computation (see Figure 3(a)), and another is the temporal derivative of the colour-saliency measure (see Figure 3(b)) The second measure is more selective than the first because it is tuned to *moving* coloured blobs at a particular scale (set by the scale of the DOG filter).

Motion triggering is the process by which the system decides whether a particular movement is worth saccading to, and tracking, or not. Naturally, this is a top-down application dependent criterion. In this case, only the movement of well-defined coloured blocks is worth tracking. The blob centre closest to the centre-of-mass of the colour-salience, temporal derivative is chosen as the most likely-location for the centre of the moving blob.


(a) Colour Selection (b) DOG filter applied to (a) (c) Zero-Crossings of (b)

(d) Spreading Activation in (c) (e) Edge Map (f) Spreading Activation in Edge-Map

Figure 4: Figure-Ground at the centre of the image

2.2 Figure-Ground at focus of attention

The Figure-Ground processes are always applied at the centre of the image, as well as any motion triggering points, if they exist.

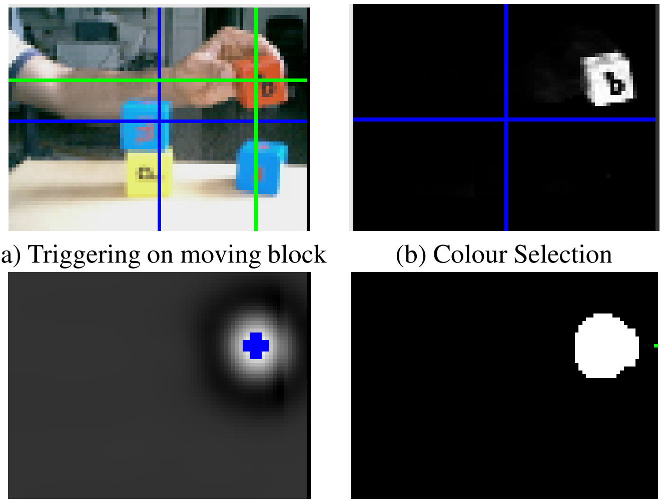
The intermediate steps in doing figure-ground are shown in Figure 4. Panel(a) shows the result of selecting the colour that's present in the fovea (a small disc around the centre of the image). The blue block at the centre and the one on the right show up clearly. Panel (b) shows the response of applying a DOG filter, on the selected colour image. The blue crosses indicate the centre of the selected blobs. Panel (c) indicates the zero-crossings of the DOG response, and (d) shows the results of a spreading-activation in the zero-crossing image from the centre. The result of the spreading activation operation enables the system to decide whether the object is well-bounded or just a background feature (in which case the activation colours the entire image). The final two panels (e)(f) show the edge-map and a similar spreading operation in the edge map. This is done simply to provide yet another source of information about the boundedness of the object.

The focus of attention need not always be only at the centre of the image. In the current example for instance, while looking at the blue block in the centre of the image the system is triggered by motion of the red block, it performs the fig-ground operations *at the triggering point* (the location of the green cross).

Figure5 shows the figure-ground operations being applied at the motion triggering point to successfully extract the red-blob. The result helps decide whether there is a well-defined block at the triggering location which is worth tracking.

2.3 Active Tracking

Once an object at the triggering location is found to be a well-defined block, the system sets the tracking model to be the characteristics of the block (colour and size) and



(a) Triggering on moving block (b) Colour Selection

(c) Blobs at triggering location (d) Confirm well-boundedness

Figure 5: Figure-Ground at motion triggering location: to decide whether the triggering object is a well-defined block

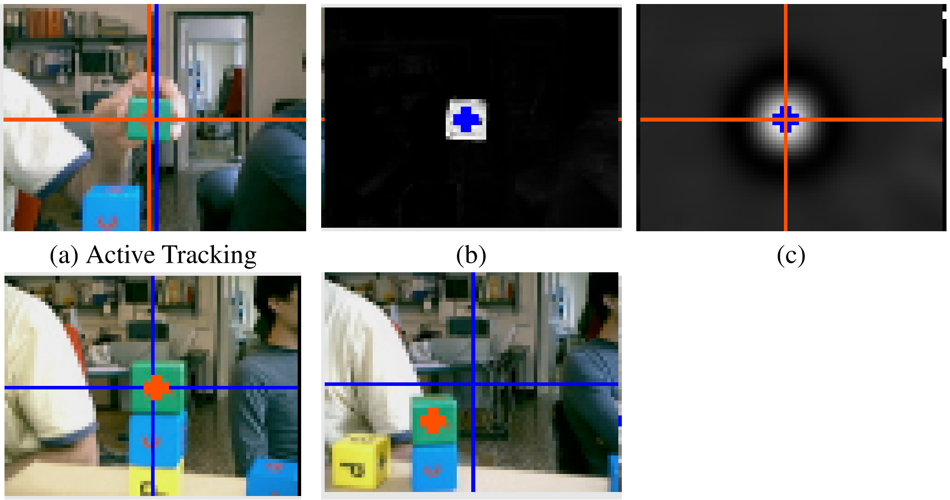
immediately saccades to that location. Now given that the system has a model of what it is supposed to be tracking, it can find the best match to that model near the centre of the image and saccade there.

Figure 6 shows such an example where the system triggers on and saccades to the green block. Panels b, c, show the colour-selection and dog filter response for the tracked green blob.

Note that the tracked object needs to only be salient to trigger the tracking, i.e. attract attention, but once the system starts using the model of the object, the tracking is robust and does not rely on the bottom-up saliency map at all.

2.4 Marker Creation and Tracking

Once the tracked object comes to rest, a marker is dropped on the object. A marker is a kind of short-term memory that binds “what” and “where” during a task, and very likely several times during the same task. Marker positions therefore have a retinal component as well as a proprioceptive component. Being important to the task, markers are always tracked, no matter what else the system is doing. The marked objects can be out of view for a while, and marker tracking resumes as soon they come back in view, this is possible only because marker’s state includes proprioceptive information. Marker tracking is therefore a combination of proprioceptive to retinal transformation followed by active tracking ,using the model of the marked object. Figure 6 panels (d) and (e) show the marker (the red-cross on the green block) being tracked across changes of block and head position. A tracked sequence is described in Figure 7.



(a) Active Tracking (b) (c)

(d) Marker Creation (e) Marker Tracking

Figure 6: Tracking, Marker creation, and Marker tracking

3 Attention

Three levels of attention are used to schedule visual resources: data driven visual attention (A1), statistically learned spatial attention (A2), and symbolically learned spatio-temporal attention (A3). Both A2 and A3 are expectation driven models which are learned from prior information. These levels of attention combine to provide *task driven attention*, since they are learned from prior observation of the typical interactions of a human with the task at hand.

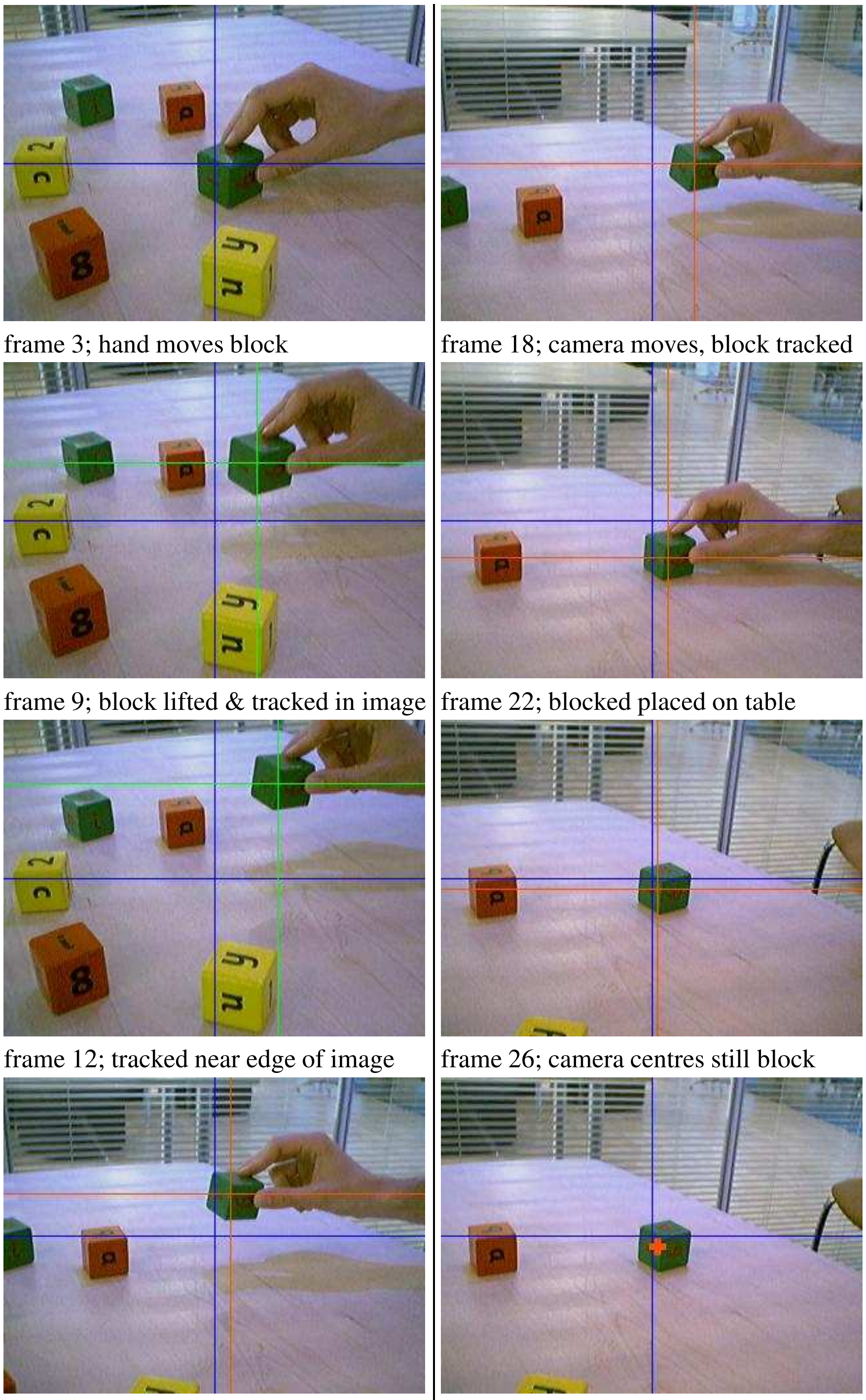
3.1 Data driven visual attention (A1)

An active vision system is essential for an embodied agent. With a restricted field of view, the camera needs to be able to move to explore the whole of the scene. This allows for the focus of attention to follow salient objects through the scene. Figure 7 shows an image sequence tracking a block with a moving camera.

A set of markers are associated with salient (interesting) objects. In this work, objects are salient if they are moving, or have moved. Attention is focused on salient objects, and the (limited number of up to seven) markers are tracked at all times. Visual attention takes highest precedence; if a salient object is being tracked, then that object will keep the focus of attention.

3.2 Statistically learned spatial attention (A2)

A probability density function in the form of a particle distribution of prototype vectors is learned using an unsupervised competitive learning neural network [2], and is essentially an online vector quantisation method. This is used to move the camera to places where objects are expected to be, and the active vision system is used to locate the object, or to report that there is no object at the location. Figure 8 shows initial prototype



frame 3; hand moves block

frame 9; block lifted & tracked in image

frame 12; tracked near edge of image

frame 14; camera moves to re-centre

frame 18; camera moves, block tracked

frame 22; blocked placed on table

frame 26; camera centres still block

frame 28; ‘marker’ dropped on block

Figure 7: Active Vision tracking system.

positions spread randomly over the camera pan-tilt space, and how the point distribution of prototype vectors has adapted to the observed data. The larger blue circle show example positions of observed objects at a single timestep.

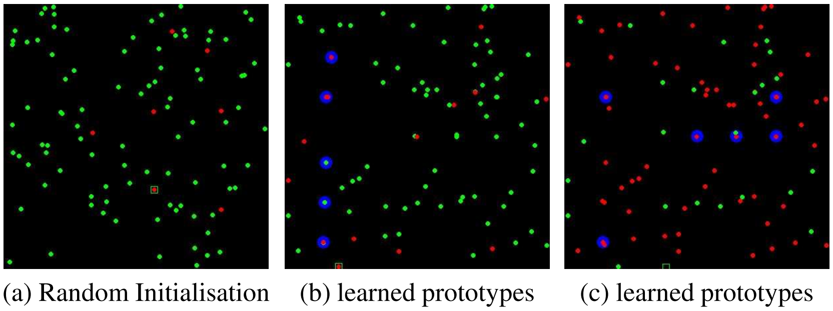

(a) Random Initialisation (b) learned prototypes (c) learned prototypes

Figure 8: Statistical learning of typical positions. Leaky neuron learning Vector Quantisation method. Blue circles show where objects have appeared. Prototype vectors have moved towards these positions.

3.3 Symbolically learned spatio-temporal attention (A3)

Spatio-temporal attention may be learned from observation of examples of activity. It can be used to as a model of where we can expect a future action to take place, given the current state of the world, or indeed previous states of the world as well. For this the modelling of spatial relationships provides a much more powerful model for expectation than modelling absolute positions.

4 Inference service

For agent behaviour generation, both continuous and symbolic models are learned. Vector Quantisation of prototypical positions are learned, which is incorporated into the A2 phase of attention discussed above. A symbolic description of the scene and activities is formed, from which spatio-temporal protocol rules are learned. These allow a cognitive agent to interact with world. They may also be used for attention A3, in which rather than being interpreted as actions that a robot should perform, they are used as a spatio-temporal attention mechanism for where to expect to see the next salient object, or motion begin.

4.1 Representation: spatial description

We are interested in the spatial relations between sets of objects in order to infer where objects should be placed to adhere to the protocol. Absolute position cannot be used for this task as it would not allow us to generalise the spatio-temporal protocol of human-object interactions that were taking place. To be independent of absolute position, a spatial calculi is needed. Many such calculi exist, though most are over-complicated for our task. We do not need a spatial description language which is as rich as say RCC-8 (which includes the ability to describe physically overlapping, or ‘fully-within’ concepts). In addition, in our applications there are many objects, and many spatial calculi are just between pairs of objects. We need to apply such a calculi to relationships between multiple objects.

A spatial relation is defined, which is independent of the absolute position of objects with respect to the image (camera pan-tilt angles), or fixed markers in the real world. This will allow for the protocol to be performed anywhere in view of the camera, and not be fixed to a rigid pre-defined grid. To do this, objects must be represented in a reference frame created from the first two objects in the scene that have moved. Figure 9 illustrates the reference frame created, forming a co-ordinate frame, with the primary axis being the vector from the centre of object A to the centre of object B, and the secondary axis being this vector rotated through 90° clockwise, to form a vector perpendicular to the first. A third object can now be described as 'left of', 'overlapping' or 'right of' object A w.r.t each axis. Figure 10 illustrates the positions that an object is in to be left, overlap of right of object A, in the reference frame formed from object A and object B. An object is said to overlap if any part of it crosses the strip of space the width of the first reference object in the direction of the axis. For each new object entering the scene, it is described in this reference frame. If an object does not have a unique description, say object D has the same description as object C, then an additional reference frame for object D is added, which is formed from object A and the object C. Using this method recursively means that a unique description is formed for each object, (and at most there are N-2 reference frames for N objects, and a small set of spatial descriptions when compared to describing every object in a reference frame made up of every other pair of objects).

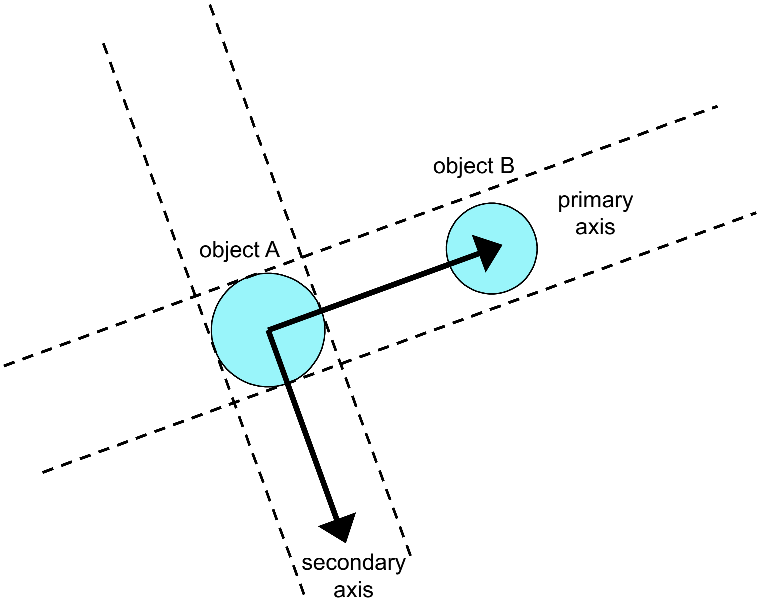
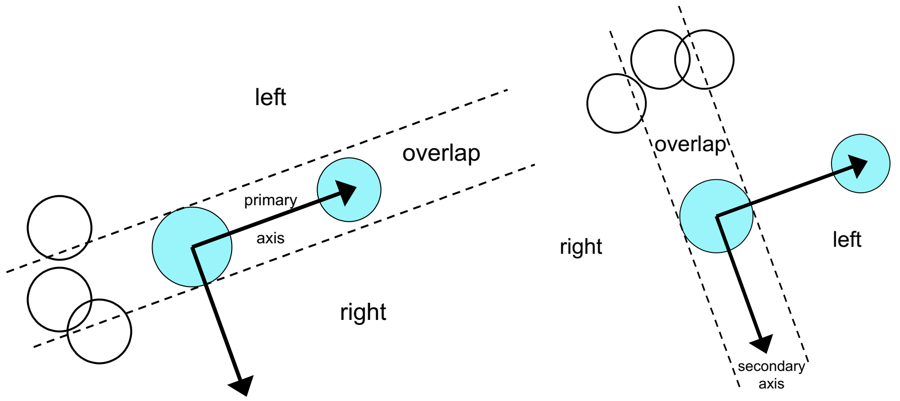


Figure 9: Spatial reference frame created from two objects.



(a) Primary Axis

(b) Secondary Axis

Figure 10: Spatial reference frame. Left, Overlap, and Right relations w.r.t (a) the primary axis and (b) the secondary axis. A relation in both axes is necessary. Example objects which "overlap" are shown by the unfilled circles.

4.2 Representation: from continuous to symbolic

Symbolic descriptions of the scene are formulated at each key-frame. A key-frame occurs at the end of a motion event. This temporal attention mechanism allows a compact representation of objects and actions in the scene. For each key-frame, we note objects and relations:

- The time and temporal successor relation
  - `time(t30).`
  - `rel(successor,t29,t30).`
- For each salient object, its existence and properties, e.g.
  - `object(obj16).`
  - `rel(state,obj16,t30).`
  - `rel(property,obj16,colour4).`
- The spatial relation of the objects with reference to two object's positions.
  - `rel(right,pri,obj17,obj15,obj18).`
  - `rel(overlap,sec,obj17,obj15,obj18).`
- The motion that has occurred. Object `obj17` is moved from where it was to a position left in the primary, and right in the secondary axis of the reference frame formed from objects `obj15` and `obj18`.
  - `rel(motion,obj17,obj15,obj18,left,pri,right,sec,t30).`

4.3 Reasoning: symbolic learning using Inductive Logic Programming

The protocols of behaviour are learned using Progol, an Inductive Logic Programming implementation [3, 4, 5]. We have previously integrated an inference engine with a visual agent to learn protocols in simple game playing scenarios [6].

The aim is to generalise the actions, or motions which take place. We wish to obtain a set of rules for deciding which block to move, and where to move it to. We may wish to say:

- move block `obj16` to a spatial position
- move a block with property `colour4` to a spatial position
- move any block to a spatial position

In order to interact with the real world, it is necessary to be able to convert back from a symbolic spatial description to a continuous pan-tilt angle description, this is done by choosing the appropriate position which is the same distance apart as the objects which form the reference frame and which fits the description. Figure 11 illustrates the possible positions in which an object would be placed (denoted by the small circles). It must be noted that when multiple reference frames are used, a unique position for the object is still obtained.

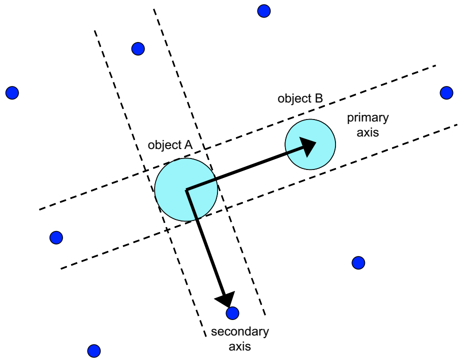


Figure 11: Spatial reference frame, with the location of 'new' positions in which to place objects given a spatial description.

5 Summary

The interaction of a spatio-temporal inference service with visual attention and active vision has been discussed. Three levels of attention prove useful for control of the camera; data driven attention for following a moving object and model based attention for spatial and spatio-temporal expectation of object locations. Once complete, the integration of an inference service which can learn, using spatial relationships, from visual observation, should produce an embodied agent able to *generalise* spatio-temporal protocols.

6 Acknowledgements

This work was funded by the European Union, as part of the CogVis project.

References

[1] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini. Learning about objects through action - initial steps towards arfificial cognition. In *Proc. IEEE International Conference on Robotics and Automation*, volume 3, pages 3140–3145, 2003.

[2] N. Johnson and D. C. Hogg. Learning the distribution of object trajectories for event recognition. *Image and Vision Computing*, 14:609–615, 1996.

[3] S. Muggleton. Inverse entailment and Progol. *New Generation Computing, Special issue on Inductive Logic Programming*, 13(3-4):245–286, 1995.

[4] S. Muggleton and J. Firth. Cprogol4.4: a tutorial introduction, 1999.

[5] S. H. Muggleton. Learning from positive data. *Machine Learning*, 2001.

[6] C. J. Needham, D. R. Magee, V. Devin, P. Santos, A. G. Cohn, and D. C. Hogg. Autonomous learning of perceptual categories and symbolic protocols from audio-visual input. In *Submitted to: British Machine Vision Conference*, 2004.

[7] Z. Pylyshyn. The role of visual indexes in spatial vision and imagery. In R. Wright, editor, *Visual Attention*. New York: Oxford University Press, 1998.

[8] C. R. Sears and Z. Pylyshyn. Multiple object tracking and attentional processing. *Canadian Journal of Experimental Psychology*, 54(1):1–14, 2000.