Visual Attention and Manipulator Control

Peter A. Sandon
Computer Science Program
Dartmouth College
Hanover, NH 03755
sandon@cs.dartmouth.edu

Abstract

One function of visual attention is as a filter that selects one region of the visual field for enhanced detection and recognition processing. A second function of attention is to provide localization information, which can be used in guiding motor activity. A visual system in which the eyes can be moved requires such localization information to guide eye movements. Furthermore, the control of arm and hand movements for object manipulation is simplified by attentional localization of the hand with respect to a fixation frame centered on the object. This paper describes this role of attention in the visual guidance of simple motor behaviors associated with unskilled object manipulation behaviors.

Introduction

It is often observed that the amount of data contained in an image is too large to be processed completely in the small fraction of a second allowed by many tasks. The obvious solution to this problem is to process only a part of the visual environment according to current task requirements. The animate vision paradigm implements this solution through the use of active control of sensors and task-dependent visual processing (Ballard:ijcai). Animate vision has been proposed as both an approach to designing computer vision systems, and as a model of human visual behavior. While the computational load is reduced when the entire image does not have to be processed, the question of what region of the image to process becomes paramount. Selective visual attention provides the mechanism for answering this question.

The term (selective visual) attention will be used to refer to a specific collection of visual subprocesses which perform the covert selection of retinal regions for further processing. This further processing may involve recognition processing of the selected region, or may involve the use of the corresponding location information for guidance of movements. It is this second aspect that will be the emphasis here, though the recognition processor is involved in this localization function, as will be discussed.

In the remainder of the paper, the mechanisms comprising an attentional visual recognition system are first discussed at a coarse level of detail. This provides a sufficient basis for describing the use of attentional localization in guiding eye and arm movements for object manipulation tasks. In particular, a touching task and a manual tracking task are used to elaborate the concepts, both of which have been implemented in a real-time robotics system to demonstrate the approach.

Attentional visual recognition

A number of computational models of attentional mechanisms have been proposed, including those of (Treisman 1988; Mozer 1988; Cave & Wolfe 1990; Sandon 1990; and Ahmad 1991). While these models differ in a number of details, and in the emphasis they place on various aspects of attentional function, they also share a number of common features which provide a sufficient basis for the current discussion. Thus, the following coarse level description of attentional visual recognition is presented for the benefit of the succeeding discussion.

The visual system consists of three components, a feature processor, an attention processor and a recognition processor (see Figure 1). The feature processor extracts a number of spatially localized features from the image. These features are extracted in parallel over the entire image, and represented retinotopically in feature maps. Though attempts have been made to identify the particular features that are extracted in

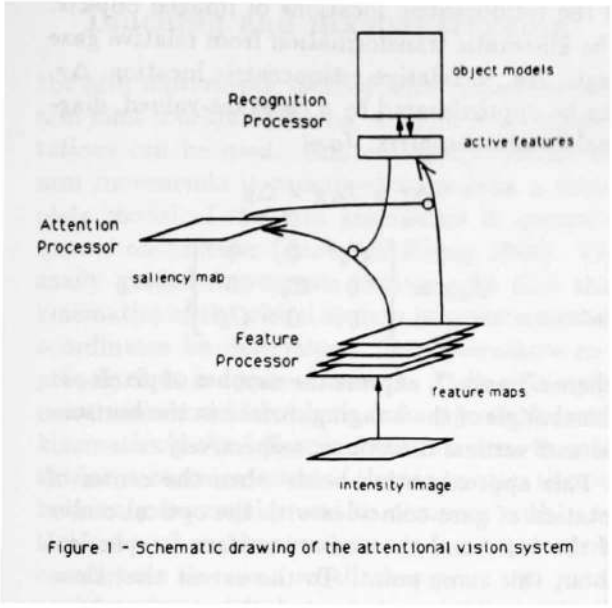

Figure 1 - Schematic drawing of the attentional vision system

human vision, this aspect is not key to the current discussion. Within each feature map, a lateral inhibition network operates on the raw feature activity, to produce contrast enhanced features. The resulting activity in each map is gated to the attention processor to a degree determined by expectancies provided by the recognition processor. Regions of activity in feature maps are also gated to the recognition processor, in this case by localized activity in the attention processor.

The recognition processor has access to a database of object models, which is indexed by feature values. Recognition is performed by having the feature processor pass image feature values to the recognition processor, which are then used to index into the object database. An object is recognized if the feature values are sufficiently close to those defining an object to satisfy some match criteria. Conversely, the recognition processor can use the defining features of an object to modify the gating of the features to the attention processor as mentioned above. We describe this use of the model data for object localization in more detail below.

The attention processor determines the region of the image whose corresponding feature values will be passed up to the recognition processor. The feature map values are combined to provide the input to a saliency map, which represents, in registration with the image, the importance of each image region to the current task. To choose a single region for processing, a selection operator is applied to the activity in the saliency map. This selection operator chooses some region of the image, whose features are then gated to the recognition processor, and whose location can be passed to motor processors.

Given these three component mechanisms, what functions might they implement? In the absence of any task-specific control of the feature map input to the attention processor, the saliency map will be sensitive to all the contrast enhanced features. The resulting saliency activity can be used to implement alerting and orienting behaviors, as well as precategorical image segmentation.

When the recognition processor activates its control of the feature map inputs to the attention processor, according to the features that characterize an object of interest to the current task, the saliency map becomes sensitive only to those specified feature maps. The selection of an active region of the saliency map in this case allows localization of the desired object in retinotopic coordinates. This location information can be used to represent spatial relations among objects, and in particular, can be used to guide motor activity, as we now discuss.

Fixation-based motor control

There has been a great deal of discussion in the literature about the appropriate frame of reference for each different aspect of visuomotor processing. While Marr, for example, emphasized the need for object-centered coordinates in representing visual information for recognition (Marr 1982), others have noted that an egocentric coordinate system would be useful when interacting with objects (Feldman 1985). Ballard argues against the egocentric representation, due to the presumed difficulty of maintaining its currency. Instead, he proposes the use of a coordinate system centered on a particular 'calibration' object (Ballard 1987).

The domain of interest here is visually guided manipulation of objects. Although it is true that the eyes, head and body may all be moving during the execution of such manipulation tasks, even a retinocentric reference frame can be effective for object localization if the spatial relations necessary to the task can be updated in a timely manner. In particular, for a binocular system, the pair of x,y coordinates representing the horizontal and vertical offsets of an object from the center of the image in each eye can be used to compute a location in a three dimensional retinocentric space.

As has been observed elsewhere (Ballard 1989), a reference frame that has particularly desirable properties is the fixation frame, which is centered on the point in space where the two optical axes of a binocular vision system intersect, and is oriented to correspond to the retinal axes and the direction of gaze. The binocular retinocentric frame is the proximal correlate of this distal fixation frame. One version of the projection of the four dimensional binocular retinal coordinates to a three dimensional space is achieved using the horizontal (h) and vertical (v) coordinates of one eye (the dominant eye), and the disparity (d) between the horizontal coordinates in the two eyes. This defines the 3-D retinocentric frame, R, in which locations are expressed as triples of the form (h,v,d). An object at the origin of this coordinate frame is at the fixation point in physical space.

The advantages in representing object location in 3-D retinocentric coordinates are that object locations can be computed quickly and maintained easily, and that the coordinate transformations required for eye movements and arm movements can be easily expressed in terms of this reference frame. The process of localizing objects in each retinal frame is mediated by the attentional mechanisms previously described. For example, to locate a particular object in one image, the recognition processor projects the feature values associated with the object to the feature processor, which differentially gates the corresponding feature maps to the attention processor. The resulting activity in the saliency map reflects the degree of match between the features defining the object and those in any particular region of the image. Selecting the most salient region corresponds to identifying the most likely location of the object in the image.

Given the two retinal locations of an object, equivalently the 3-D retinocentric coordinates, the guidance of eye and hand movements toward the object is relatively straightforward. For eye movements, the motor frame is defined by the gaze angles of the two cameras. Analogous to the 3-D retinocentric frame, the appropriate gaze angle frame, G, for a pair of horizontally offset, fixating eyes, is a 3-D reference frame consisting of the yaw angle (θ) and pitch angle (ϕ) of the dominant eye, and the yaw angle disparity (ψ) between the two eyes.

Eye movements are defined relative to the current gaze, and result in a relative displacement of the retinocentric locations of imaged objects. The kinematic transformation from relative gaze angle, Δg, to relative retinocentric location, Δr, can be approximated by a constant-valued, diagonal Jacobian matrix, J_{GR}:

$$\Delta r = J_{GR} \times \Delta g$$

$$J_{GR} = \begin{bmatrix} C_h & 0 & 0 \\ 0 & C_v & 0 \\ 0 & 0 & C_h \end{bmatrix}$$

where C_h and C_v express the number of pixels per visual angle of the imaging surface in the horizontal and vertical directions, respectively.

This approximation holds when the center of rotation of gaze coincides with the optical center of the lens, and the sensory surface is spherical about this same point. To the extent that these two assumptions are violated, the constant function kinematics will be less accurate, though for small gaze angles and limited depth of field, the accuracy will remain high.

The process of establishing a new fixation point is as follows. If the desired action is to fixate a particular object, the object is first localized in each image as described previously. The locations in the two images are used to compute a location, r, in the retinocentric frame, R. Since the desired location is at the origin in R, the vector -r represents the relative movement in R. This vector is passed to the eye movement control system, which computes the transformation from R to G as:

$$\Delta g = J_{GR}^{-1} \times -r$$

The computed gaze angles are used to direct a saccadic movement of the eyes to the new fixation point.

In the absence of having a particular object specified as the target of fixation, the process remains the same, except that the feature maps are gated to the attention processor according to some default weighting of the individual maps, corresponding to the relative importance of each feature for alerting purposes.

This scheme can be extended to smooth pursuit eye movements by performing an additional filtering step on a sequence of gaze angle values that are obtained by successive executions of the above procedure. To maintain accurate pursuit, a predictive filter such as a proportional-integral-derivative (PID) filter can be used to adjust gaze velocities (Dorf 1986).

Touching and manual tracking

For arm movements, defined with respect to the arm joint coordinate frame, A, analogous computations can be used. Conventionally, control of arm movements is presumed to require a complete model of the arm kinematics in environmental coordinates (Brown & Rimey 1988). Visually guided movements then require that the kinematics of the visual system in environmental coordinates be determined. An alternative approach, that is applicable to the kinds of simple movements considered here, is to express the arm kinematics in the 3-D retinocentric frame. In particular, a representation of the kinematics that is both easy to acquire and to compute with is a local one, where the small change in retinocentric coordinates due to a small change in arm joint positions is used to represent a constant-valued kinematics in that particular region of joint-gaze space (Mel 1989). That is, for a particular joint-gaze configuration, the change in retinocentric coordinates, Δr, for a given change in arm joint positions, Δa, is given by:

$$\Delta r = \hat{J}_{AR} \times \Delta a$$

where $\hat{J}_{AR}$ is the Jacobian evaluated at the particular joint-gaze configuration.

One way to represent the complete kinematics is as a collection of evaluated Jacobian matrices indexed by joint-gaze coordinates in a lookup table. These matrices can be acquired through a calibration procedure prior to use, or through an adaptive process during movement execution. This has advantages for acquisition and for representation of arbitrary relations. Alternately, a representation of the Jacobian terms as low-order functions of joint-gaze space is more efficient and provides better generalization during acquisition when the relations being represented are smooth. The direct kinematic equation above is used for acquisition of the kinematic parameters, while the inverse Jacobian is used for control.

Touching

Perhaps the simplest object manipulation behavior is touching, that is, using arm movements to bring the hand into proximity with some object of interest. Given the previously described attentional mechanism for locating objects in R, and kinematic models for transforming between R and G, and between R and A, the touching task can be accomplished as follows:

TOUCH (object):

$r = Attend$(object)	;locate object in R
$\Delta g = J_{GR}^{-1} \times -r$	;saccade to the object
$r = Attend$(hand)	;locate hand in R
$\Delta a = \hat{J}_{AR}^{-1} \times -r$	;move hand to object

Due to the use of local kinematics, a given move will be inaccurate to the degree that the new joint state is far from the initial one. This approach is appropriate, therefore, when a lack of real-time constraints allows for the use of one or more small compensatory movements to be used to achieve the desired accuracy.

Notice the minimal need for representation of spatial relations in this process. Attention is first used to locate the object of interest. This location information is represented in the state of the selection process, which is transmitted to the eye movement control system. Once the eyes have been moved, the location of the object is implicit in the gaze angles of the eyes, and the attention processor need not maintain that location (which is now out of date in any case). Attention is now used to locate the hand, and the selection process represents the location for the sake of the arm movement control system. There is no need to maintain location information across movements for this simple task, because it can be easily reacquired by repeating the sequence.

Tracking

A relatively simple extension of the touching behavior allows a moving object to be manually tracked. We will use the term pursuit to refer to eye movements that maintain fixation on a moving object, and manual tracking, or simply tracking, to refer to arm movements that maintain proximity of the hand to a moving object. Although the tracking behavior by itself is not one that is commonly executed, it is a necessary component of tasks that require moving objects to be grasped, and a precursor to tasks that require interception of moving objects, such as catching and hitting. More importantly for the present purposes, the tracking behavior demonstrates the use of the attentional mechanism as a shared resource for the concurrent control of the eye and arm motor systems.

The tracking task could be accomplished by simply executing the touching behavior in an iter-

ated loop. However, this yields a sequence of discrete movements for the eyes and the arm, rather than the smooth movements that might be desired. The required modification is straightforward. The attentional processor toggles back and forth to locate first the object, then the hand, as in the touch procedure. The locations that are supplied to the motor control processes are then transformed by a predictive filter. The output of the filter is used to control the gaze and arm joint velocities, such that the object being tracked is maintained at the fixation point, and the hand is maintained close to the object:

TRACK (object):
repeat

$\mathbf{r} = \mathit{Attend}$(object)	;locate object in R
$\Delta g = J_{GR}^{-1} \times -\mathbf{r}$	;desired gaze change
$\Delta \dot{g} = PID(\Delta g)$	;smooth gaze adjust
$\mathbf{r} = \mathit{Attend}$(hand)	;locate hand in R
$\Delta a = \hat{J}_{AR}^{-1} \times -\mathbf{r}$	;desired arm change
$\Delta \dot{a} = PID(\Delta a)$	;smooth arm adjust

An implementation of the saccade, pursuit, touching and tracking behaviors just described has been developed for a binocular camera and robotic arm system. The vision system consists of a pair of cameras mounted on a motorized pan-tilt platform, and a Datacube Maxvideo image processing system. The arm is a PUMA 761 six degree-of-freedom arm. A SUN4 workstation runs the control program and mediates communication between the image processing, eye motor control and arm motor control systems.

The features used for defining objects are based on image intensity, edge orientation and edge ratio magnitude. The object of interest is attached to a slowly revolving platform placed within the workspace of the arm. The pursuit behavior has a .4s cycle time, and generates a smooth gaze trajectory that lags the object by up to a degree in each dimension. The tracking behavior has a 1.25s cycle time, and generates discrete arm movements, due to a lack of velocity control in the current arm controller interface. These movements also lag the object movement, and exhibit an appreciable rms error from the expected trajectory, that is four times greater (48mm vs 12mm) in the direction parallel to the line of sight than in the directions perpendicular to the line of sight.

Further details are presented in (Sandon 1992).

Concluding remarks

Although a great deal of consideration has been given to the mechanisms of attention, much less work has addressed the function of attention in everyday visuomotor behavior. This paper describes, and the briefly presented implementation results demonstrate, a computationally simple approach to visual guidance of eyes and arms based on attentional localization and local kinematics. The minimal representation used in the approach has advantages in computational efficiency, both for acquiring and for maintaining a current model of the external world. In addition, minimal representations exhibit advantages in adaptive systems, since the credit assignment problem is reduced (Whitehead & Ballard 1990).

As stated, this approach to object manipulation applies to servo-controlled movements, in which visual feedback is used to repeatedly adjust an eye or arm movement. This is an appropriate model for unskilled behavior, and corresponds to a situation in which the kinematic and dynamic models of the motor systems are not well characterized. While more complete and accurate models are required for modelling skilled movements and for tasks having significant real-time constraints, it seems reasonable to assume that such models are preceded by the approximate ones discussed here. More accurate models are then acquired using the errors that occur while performing these simpler behaviors.

While it may seem intuitive that covert attention should be used to guide overt eye movements, the precise relation between the two systems is not yet clear. On the one hand, Remington found that the enhanced processing associated with attention preceded saccadic eye movements that were initiated by a stimulus onset in the retinal target position (Remington 1980). This provides evidence that attention is being used to guide the eye movement. In addition, there is evidence that one component of saccadic latency is the time needed for attention to disengage prior to localizing a target to be fixated (Fischer & Breitmeyer 1987). However, Remington also found that for eye movements initiated by a central cue indicating the desired direction of movement, attention followed the eye movement to the target position, indicating that saccadic guidance was provided by some other source. As for the guidance of arm movements, there is evidence that eye movements play a part (Ballard, et. al. 1991), but the role of attention is not known.

How does this approach extend to more complex tasks? The introduction of real-time constraints has already been mentioned. These require accurate ballistic movements, which in turn require more accurate kinematic and dynamic models. As previously discussed, these models can be developed during the execution of the simpler behaviors described here. When the task involves the manipulation of additional objects, attention must be shared among the objects to maintain localization information. Furthermore, an explicit short term representation of objects will likely be necessary, in order to maintain continuity of object characteristics, and to predict future object location for guiding the selection process.

Finally, for more complex interactions with objects, in particular, for grasping them, hand movements must be controlled in addition to eye and arm movements. Grasping behaviors require not only localization of an object, but an estimate of object pose. In many cases, scale and major axis orientation information are sufficient for the determination of an appropriate hand configuration for grasping. For more complex objects, detailed pose must be determined. While desirable features for localizing an object are those that do not depend on viewpoint, the features needed to determine pose are those that are viewpoint dependent. In addition, the likely role for attention in detailed pose estimation is in localizing the components of objects to represent the spatial interrelations among parts.

Acknowledgements. This work was done while the author was visiting the University of Rochester. Thanks to Dana Ballard, the Vision Lab research group, and the CVS Visual Attention reading group for many useful discussions. This research was supported by NSF Grant No. IRI-90108999.

References

Ahmad, S. 1991. VISIT: An efficient computational model of human visual attention, Tech. Report, TR-91-049, International Computer Science Institute.

Ballard, D. H. 1989. Reference frames for animate vision. In Proc. of the Eleventh International Joint Conference on Artificial Intelligence, 1635-1641, IJCAI, Inc.

Ballard, D. H. 1987. Eye movements and spatial cognition, Tech. Report, 218, Dept. of Computer Science, Univ. of Rochester.

Ballard, D. H.; Hayhoe, M.; and Li, F. 1991. Hand-eye coordination during sequential tasks. Forthcoming.

Brown, C. M.; and Rimey, R. D. 1988. Coordinates, conversions, and kinematics for the Rochester Robotics Lab, Tech. Report, 259, Dept. of Computer Science, Univ. of Rochester.

Cave K. R.; and Wolfe J. M., 1990. Modeling the role of parallel processing in visual search. *Cognitive Psychology* 22:225-271.

Dorf, R. C. 1986. *Modern Control Systems*, Reading, Mass.: Addison-Wesley.

Feldman, J. A. 1985. Four frames suffice: A provisional model of vision and space. *Behavioral and Brain Sciences* 8:265-289.

Fischer, B.; and Breitmeyer, B., 1987. Mechanisms of visual attention revealed by saccadic eye movements. *Neuropsychologia* 25:73-83.

Marr, D. 1982. *Vision*. San Francisco: Freeman.

Mel, B. W. 1989. MURPHY: A neurally-inspired connectionist approach to learning and performance in vision-based robot motion planning, Tech. Report, CCSR-89-17A, Univ. of Illinois.

Mozer, M. C. 1988. A connectionist model of selective attention in visual perception. In Proc. of the Tenth Conference Cognitive Science Society, 195-201, Hillsdale, NJ.: Lawrence Erlbaum.

Remington, R. W. 1980. Attention and saccadic eye movements *J. Exp. Psych.: HPP* 6:726-744.

Sandon, P. A. 1992. Visually guided touching and manual tracking, Tech. Report, 412, Dept. of Computer Science, Univ. of Rochester.

Sandon, P. A. 1990. Simulating visual attention. *J. Cognitive Neuroscience* 2:213-231.

Treisman, A. 1988. Features and objects: The fourteenth Bartlett memorial lecture. *Quarterly J. Experimental Psychology* 40A:201-237.

Whitehead, S. D.; and Ballard, D. H., 1990. Learning to perceive and act, Tech. Report, 331, Dept. of Computer Science, Univ. of Rochester.