A Survey of Selective Fixation Control for Machine Vision

A. Lynn Abbott

A number of methods can be used for automatically directing camera gaze to new visual targets. Recent research has shown that computational advantages may be realized for computer vision problems when cameras are capable of fixating an object of interest. Here, we are not concerned with methods of acquiring or maintaining fixation, but with a dual problem: how to select new points of interest to be fixated. The emphasis is not on physical aspects of controlling camera movements, but on quantitative methods used in the selection of new targets for fixation. Because many of these methods are motivated by human visual perception, a brief summary of relevant research in human vision is given here as well. This is followed by a discussion of computational models for selective fixation control, most of which have begun to emerge only in recent years.

Machine Vision Research

Traditionally, research in machine vision has concentrated on the analysis of passively acquired images. This contrasts with human perception, for which visual sensing is an active process, tightly integrated with behavior and strongly motivated by particular goals. For example, a person explores a room or recognizes a face by successively looking at different points of interest until the task is completed; the observer does not merely obtain a single glimpse and then retreat to consider what has just been seen. In essence, the active observer "interrogates" the environment through a sequence of sensory probes so that new information is obtained and internal representations are updated.

The process of "looking" at a single point in the three-dimensional (3D) scene is known as fixating, and the location at which the eyes (or, equivalently, cameras) are aimed is known as the fixation point. Recent research suggests that computational advantages may be realized for vision problems when cameras are capable of fixating an object of interest [6], [7], [34]. An important consideration in the example above is that fixation locations are not chosen at random, nor are they rigidly planned in advance. Instead, active perception [5] implies that fixation points must be selected as a result of new visual information and based on an evolving internal model of the environment. We refer to the process of identifying and selecting new visual targets as selective fixation control. (See Fig. 1.) This is sometimes known as the "view-selection" or "next-look" problem. Fundamentally, selective fixation facilitates the dynamic, cooperative interleaving of image acquisition with scene analysis.

In general, a fixation point may lie on an object which is in motion with respect to the observer. In order to maintain attention at this point, some mechanism is needed which tracks the object either by physically shifting the gaze direction or by logically shifting a window in an image. In either case, fixation of a moving object implies smooth tracking movements which may be expressed and analyzed using the more traditional tools of control theory (for example, see [13]). In contrast, selective fixation implies fast, discontinuous attentional shifts to new targets within the visual environment. The selection of new targets is a higher-level problem which decides what feature or object is to be fixated or tracked. This is not so easily represented within the traditional control framework. Visual tracking and selective fixation control are distinct and complementary capabilities, and both are needed for visual sensing by an active observer.

This paper describes several methods that have recently been developed for selecting new visual targets in machine vision systems. Interest in this topic has paralleled the technological advances which now permit the relatively inexpensive computational control of camera movements and image acquisition. This paper is particularly concerned with the identification and integration of criteria which underlie the target-selection process. The choice of a single visual target often implies the resolution of conflict among several competing criteria. Ultimately, selective fixation facilitates the overall visual sensing strategy for an autonomous agent.

Because many of these methods are motivated by human perception, the next section summarizes criteria used in human vision for selecting new fixation locations. The paper then describes how these and other criteria have been adapted for selective fixation in machine-vision systems. The last sections contain a brief analysis and comparison of these methods.

Selective Fixation in Biological Vision

Eye Movements

Human eye movements are integral to visual perception. When confronted with a new scene, the visual system very quickly organizes and executes a sequence of eye movements which are used to build an overall perception of the scene. These movements are characterized by brief periods of fixation, punctuated by rapid movements to new fixation points.

During the fixation periods, the visual system relies directly on visual information as it attempts to stabilize the retinal image. For moving objects, this involves "smooth pursuit" movements to track the object of interest. In contrast, rapid (saccadic) eye movements are discontinuous jumps which seem to transfer fixation to a new locations instantaneously. Saccades have been called "preprogrammed" or "ballistic" to emphasize that most processing of the retinal image is sup-

A preliminary version of this paper was presented at the IEEE International Conference on Systems, Man, and Cybernetics, Charlottesville, VA, October 13-16, 1991. The author is with the Bradley Department of Electrical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.
pressed during the movements. In other words, a visual target must be chosen in advance, before the saccade occurs. The visual system then computes the rotation angle needed to transfer the direction of gaze to the vicinity of the chosen target, and executes the movement. Since saccadic movements typically occur 3 or 4 times per second, fast mechanisms are required in the selection of new fixation targets.

Saccades are especially important for human vision because of the nonuniform resolution of the retina. Eye movements are used to bring the image of a desired object onto the fovea, where resolution is highest. This serves to concentrate attention near the image centers, facilitates stereopals, and implies that visual targets are often selected from the low-resolution periphery. It is a curious fact that the world is perceived as stable and continuous, despite nonuniform resolution, frequent saccades, and viewer motion.

**Selective Attention without Eye Movements**

Eye movements are not the only means of directing attention to new regions in a scene. As early as 1850, researchers concluded that it is possible to shift attention from one location in the visual field to another without eye movements [21]. These attentional shifts occur in discrete jumps across the image, can be voluntarily directed, and for some tasks require time which is independent of distance on the retinal image [33], [25]. Furthermore, in some cases selective attention without eye movements has been shown to be instrumental in recognizing objects. (Peterson and Gibson demonstrate this for modified Necker cubes [31].)

The shape and extent of the attentional areas in the image can vary. Some have suggested that the “focus of attention” can change from a uniform distribution over the entire visual field to a small, concentrated region. Such selectivity has been likened to a spotlight or zoom lens [17]. Attention can also be directed to stimuli in noncontiguous regions. For example, it is possible to concentrate simultaneously on several moving objects against a stationary background [29].

**Factors which Guide Shifts in Attention**

Whether or not eye movements are involved, attentional shifts must be directed. We use the term selective fixation for either case. These shifts may depend on both visual and nonvisual factors. The following paragraphs list several salient criteria and summarize studies of eye movement behavior and attentional shifts. The conclusions listed are representative of human visual behavior, and are not necessarily true in all situations. (This expands on a similar discussion in [1].)

**Low-level visual stimuli.** Fundamentally, any detectable feature can be used to guide attentional shifts. For example, color can be an important factor in visual search [14]. Other apparent attractions are high-contrast regions and image areas of high spatial frequency.

**Higher-level visual features.** When examining pictures or photographs, saccades are often directed to areas of “informative detail” [28]. This often involves the recognition of such high-level image components as human faces. For simpler objects, such as polygons, eye fixations tend to concentrate near vertices [8]. When symmetry is present in 2D displays, subjects tend to concentrate fixations along the axes of symmetry [27].

**Sudden change.** The sudden appearance or disappearance of a peripheral stimulus often leads to a saccadic movement toward that location [18]. However, it has been shown that transient visual stimuli do not attract attention when outside the focus of attention [36].

**Motion.** It has long been known that movement in the periphery provides a strong incentive for eye movements. It has been demonstrated that the visual system can distinguish particular forms of motion characteristics and can facilitate the direction of attention to those regions of the visual field [29]. For example, attention can be directed to several moving items when stationary stimuli are also present, even if object motion is not correlated. Also, attention can be directed to stimuli moving in one direction even if other objects are present and moving in another direction.

**Proximity of stimulus.** When several potential targets are present in the visual field, those which are closer to the fovea are more likely to be selected for fixation [18].

**Direction of stimulus.** Levy-Schoen describes a study in which upward eye movements were preferred over downward movements [26]. Furthermore, subjects often selected visual targets in a “centrifugal” order, relative to an initial fixation point. This suggests that proximity to the original fixation location is an important criterion, especially since this tends to reduce the total distance “traveled” by the gaze.

**Dependence on high-level goal.** Although some early researchers assumed that visual scanning was random, Yarbus clearly demonstrated that scanning is highly goal-dependent [42]. (A quantitative discussion appears in [16].) He asked subjects to examine a picture, and recorded eye movements as the subjects attempted different simple tasks, such as estimating the ages of people or locations or objects shown in the picture. The resulting scan patterns were remarkably different for each task.

**Supplement to internal representations.** O’Regan and Levy-Schoen suggest that eye movements occur to provide additional details that are not maintained internally [33]. They offer the metaphor that the visual world serves as an “external memory buffer” which can be accessed using eye movements. Since humans appear not to maintain an internal spatial map with high accuracy [20], visual probes can provide current high-resolution information to supplement coarse internal representations.

**Selective Fixation for MachineVision**

Fixation control in machine vision systems has been motivated by studies of human visual behavior, but additional criteria have been identified on purely computational grounds. This section describes several research efforts of selective fixation and attention. Each of the systems considered must address the following question: For a given scene and sensing goal, what is the strategy for selecting new fixation points? For several of these systems, this involves the movement of one or more cameras so that the new visual targets are fixated. In a few cases the cameras do not move, but the system attends to the new target in the image by computational or windowing methods. The systems are grouped roughly according to several broad sensing goals, and we are particularly interested in the factors used in the selection of new fixation locations. Where appropriate, we also consider the methods used for resolving conflicts among competing visual targets. Some of these methods have been tested with physical implementations; some have been tested by simulation; and some of the methods are strictly theoretical. All of the methods fall under the category of active vision [4], [34].

**Visual Search and Object Detection**

Visual search is the process of seeking instances of target objects within a given scene. Typically, a great deal of visual data must be considered and classified as either “target” or “background” information. Some of the methods considered here are quite general, in that they...
are based on the integration and resolution of feature types that can be arbitrarily specified.

Koch and Ullman [23] describe a mechanism for shifting attention across the visual field based on an abstract measure of "saliency." The method assumes that any number of elementary features (such as color or orientation) are available from computations performed in parallel across the entire image. For each image location, a quantitative measure of saliency is obtained through a combination of the low-level feature values at that location. (The method of combination is not given explicitly.) The maximum value in the resulting saliency map is then chosen as the next visual target. Attentional shifts can occur as the image changes, or by dynamically altering the relative weightings of different feature types. The discussion is theoretical, and is strongly motivated by physiological concerns.

Clark and Ferrier [12] describe a physical implementation which uses the above method to direct the movements of stereo cameras. A two-level control is used, in which the lower level is a traditional feedback control system, modeled after the human oculomotor system and capable of visual tracking. The higher level in this hierarchy is based on a realization of Koch and Ullman's saliency map to select new targets for fixation. For each image, a saliency map is computed as the weighted sum of elementary feature values for each pixel in that image. The maximum values in these maps determine the locations for fixing the respective cameras. The weights can be dynamically altered to force shifts in attention. This system is generally applicable since different weightings and elementary feature types can be implemented as needed.

Burt [10] has considered hierarchical approaches to the selection of visual targets. The motivation is to restrict attention so that scarce computational resources are appropriately and dynamically allocated. Pyramid-based implementations have been tested for locating areas in the image which are more likely to contain information pertinent to the task at hand. As an example of a surveillance system, a method is described in which difference images are obtained for successive video frames from a stationary camera. Nonzero values result only where temporal changes are present, and represent candidate targets for fixation. To reduce processing requirements, a Laplacian pyramid is obtained for the difference image. If the finest resolution image is squared and integrated, motion is then apparent at the coarsest level of resolution, simplifying the overall detection process. After detecting the object at the coarsest level, a foveation procedure can occur during which an image window passes through the resolution levels until centered on the moving object at highest resolution. The object may then be tracked for further processing.

Burt also suggests a format for object recognition, in which image features are maintained at different resolution levels within a tree structure. Visual recognition often involves the detection of larger features first at coarser resolution, followed by a search for patterns with progressively more detail. In general, features with lower resolution are chosen from this border by minimizing the following objective function:

\[ E = a_1 \left( \| p - p_{\text{CAM}} \| \right) + a_2 A \left( p, p_{\text{PROF}} \right) + a_3 A \left( p, p_0 \right) \]

In this equation, \( p \) is a candidate fixation target, \( \| \cdot \| \) represents 3D Euclidean distance, and \( A \) is a function which maps two 3D points to the angle between the lines of projection for those points. The points \( p_{\text{CAM}}, p_{\text{PROF}}, \) and \( p_0 \) represent the current location of the left camera, the current fixation point, and the initial fixation point, respectively. The constants \( a \) permit relative weightings among the three terms.

This method is therefore based solely on proximity criteria, rather than on visual stimuli. The first term favors scene points which lie near the cameras; this helps the system initially to avoid distant objects and causes the scan to proceed from near to more distant locations. The

August 1992 27
reason for this is to avoid problems involving occlusion boundaries. The second term represents a bias for new visual targets that are near the previous fixation point; this tends to prevent large camera movements during the mapping of a single object. The final term favors scene locations that lie near the original fixation point; this tends to cause a global depth map to evolve radially outward from that initial location. After choosing the target, the system automatically aims both cameras and adjusts lens focus remotely to estimate depth accurately in the vicinity of the new fixation location. The system is intended for building a surface map of a single piecewise-smooth object, which may be large relative to the visual field. New fixations are expected to smoothly extend the regions of estimated depth in the global map into unknown portions of the scene.

Das and Ahuja [15] extend the above system to accommodate scenes with multiple objects. They use the same method for selecting targets, except that the global map now may contain coarse depth information in addition to high-accuracy stereo estimates. Because this system automatically focuses the fixation point, peripheral image areas may be defocused. The coarse estimates are obtained from these peripheral areas. After a target is chosen, the cameras are not so simply fixated at the new location since the distance estimate for a defocused target may not be very accurate. The system therefore performs a “homing” sequence, in which focus is gradually changed to guide the two cameras to a new fixation location in the vicinity of the selected target.

Shmuel and Werman [35] describe a method for constructing a global surface map from a sequence of monocular images. Each successive camera position is chosen with the goal of providing new information for optimal refinement of the surface map. The camera is constrained to translate in the plane perpendicular to its optical axis by a small (predefined) amount; this leaves one degree of freedom (direction of translation) for selecting a new camera position. After each image is obtained, the surface map is updated using motion disparities and a new camera direction is chosen by optimizing an objective function which is based on the expected reduction of uncertainty in the map.

Scene depth is maintained as a dense map, where each value represents the reciprocal of depth. A Gauss-Markov discrete time model is used

\[ M_t = H_t s + v_t \]

where the state \( s \) represents the actual (reciprocal) depth values to be determined, and \( M_t \) is the estimate of \( s \) at time \( t \). The vector \( v_t \) represents additive white Gaussian noise with zero mean and known covariance \( R \), and \( H_t \) is a linear transformation matrix. After each new image is obtained, a Kalman filter is used to update \( m_t \) and a corresponding uncertainty map. Intuitively, new directions of translation will tend to be perpendicular to image contours, since this yields image features which provide more accurate disparity measurements.

Whait and Ferrie [40] describe a system which characterizes the 3D geometry of objects by estimating superellipsoid parameters from range data. The system performs a least-squares fit of a superellipsoid to a region, using radial distance from the model surface to each data point as an error measure. A problem with this approach is that a significant subset of the parameter space may correspond very nearly to the globally optimum fit. These ambiguities may result from noise, or simply because many different models will fit the data perfectly well. The authors use a covariance matrix (a product of the fitting procedure) to represent this lack of uniqueness. This represents a measure of uncertainty, and can be used to derive a confidence interval around the true model parameters. New viewpoints are chosen in succession so that this interval is optimally improved. In some cases, it is not sufficient merely to move the viewpoint to the occluded side of an object. This method has been tested using range data of a wooden mannequin.

Object Recognition

Systems which can recognize objects in real-world situations must contend with a large number of object models and with an enormous amount of visual data. Selective search algorithms suggesting regions of the visual field where important information exists, thereby reducing computational complexity.

Grimson [19] describes a method for recognizing objects which are modeled as polyhedra. It is assumed that the sensor, possibly a camera, is capable of estimating the directions of surface normals. After the system obtains a sparse set of initial surface-normal measurements, a recognition system generates a small set of candidate objects with associated pose information. In order to select the correct interpretation from this set, the system needs to select new, optimal sensor locations to obtain additional surface information. Assuming that the sensor is constrained to translate within a plane, new view positions are chosen so that the resulting lines of sight will fall on the desired surfaces which best disambiguate among the candidate objects. In order to determine these lines of sight, all visible faces for each candidate model are projected onto the plane in which the sensor translates; planes that have not yet been sensed are chosen as targets. This method has been tested by simulation.

Browse and Rodrigues [9] consider the problem of selecting new candidates for eye-camera fixation within graded-resolution images. These are images having resolution which decreases with distance from the image center, similar to the human retina.) A process called “propagation” is defined which integrates image information at the different resolution levels. Information from the periphery is examined to predict locations which, if fixated, should have the highest probability of enhancing visual analysis. The method is not limited to object recognition. It has been tested in the context of recognizing line drawings the human body. A graded-resolution line drawing is presented to the system, along with knowledge of the organization of the human body in the form of graph-based models. Knowledge of the connectivity of different body parts (upper body to arm, arm to hand, etc.) is used to assess different possible interpretations of the image, and to suggest new locations to be fixated.

Califano et al. [11], describe a system in which an image is processed simultaneously at two different levels of resolution. Processing occurs at low resolution over the entire image, and higher-resolution computation is performed for a small window (analogous to the retinal fovea.) They describe a method for controlling movements of this high-resolution window within the image to assist in object recognition. Although they consider this in the context of analyzing range images, these methods could also apply to intensity images. The system uses 2D texture and 3D curve as features.

A set of independent “control operators” is defined, each of which specifies a candidate target location in the image for the foveal window. One of the fixation targets is chosen through a conflict-resolution mechanism. These operators can be model driven (by processing models of objects to be recognized) or data driven (primarily dependent on sensor data). Three control operators are described. The first is a model contention operator which detects conflicts among object hypotheses. For example, if two objects appear identical at the coarse level, this operator suggests a fixation location at which the objects would differ at higher resolution. The second is an unexplained region operator,
which detects large, unexplained regions in the image. These regions are prioritized, and the centroid of the region with highest priority is chosen as the candidate target. The final one described is an explore operator, which selects a new fixation target within an unexplored portion of the image based on a simple, predefined scanning scheme. Conflict resolution is performed by assigning heuristic weights to each operator and then selecting the one with the largest resulting value. This system can accommodate a wide range of visual tasks, including surveillance and object location. The system has been tested using range images, and successfully distinguishes between a golf ball and a ping-pong ball, and between gears with different numbers of teeth.

**Autonomous Navigation**

Visual navigation, in general, is a very difficult since it may depend on locating and recognizing objects along a desired path. In a real-world environment, the autonomous agent must identify obstacles and destinations and must use these to plan an appropriate path.

Ishiguro et al. [22], describe a framework in which a robot must use vision to stay along a predetermined path. This path lies on a planar surface, and a single camera is attached to the side of the robot. As the robot moves forward, the camera fixates an object feature and tracks the feature by rotating about its vertical axis. A method is given for generating a local depth map about the fixation point by using knowledge of camera rotation. As the robot continues to move, new objects come into view and the camera must acquire new fixation targets. The estimation of local depth maps has been tested with an actual robot which moves in linear and circular paths. The selection of new fixation points is still under investigation.

Xie [41] presents two heuristic methods for planning viewpoints in a known 2D environment. The world is assumed to contain opaque polygons which are rigid and stationary. The goal is to select a sequence of camera locations which together will cause all object faces to come into view. Using a generalized Delaunay triangulation, heuristics are used to merge triangles into star polygons which partition or cover the navigable space. The resulting algorithms are $O(N \log N)$. An application of this method might be autonomous patrol, in which a robot verifies that the environment is as expected. A method is also given for exploring a similar, unknown environment. Each successive viewpoint is chosen within an empty portion of the area which has already been explored. The method assumes that location information is obtained for each object face as it comes into view. From each view, a fan or star polygon is obtained. An $O(N_2 \log N)$ algorithm is given for selecting the next viewpoint. Both of these methods are illustrated with examples, but no simulations or implementations are described. The resulting paths are not necessarily optimal, but the method offers adequate solutions in a reasonable amount of time.

Rimey and Brown [32] consider the use of augmented hidden Markov models (HMM) for learning and generating sequences of camera fixations. Roughly speaking, a HMM is a finite state machine with a probability assigned to each state transition. This may be represented as a directed graph with a weight associated with each arc of the graph. It is possible to "train" a HMM by presenting a sequence of state transitions and modifying arc weights to reflect the frequencies of state transitions. An augmented hidden Markov model (AHMM) is defined to permit refinement of these weights over time, based on image data.

Rimey and Brown describe the use of two HMMs to guide the selection of fixation points for object recognition. One graph contains information about the order in which a set of visual features should be fixated for a particular object; the second graph contains knowledge about where the features exist in the scene. Experiments have involved the use of camera movements to recognize a line drawing of a human face. The imaging system produced a high-resolution central image, with an accompanying low-resolution periphery. In recognizing a face, it is common for a human observer to fixate such features as eyes, nose,
and mouth in sequence. The HMMs used here guided a similar examination of the line drawing by selecting a particular feature type, and then hypothesizing the best location to be fixated for this.

**Discussion**

In addition to the physiological criteria identified in the second section, some of the computational methods in the last section have employed additional factors in selecting visual targets. These include:

1. establishment and use of multiple object-centered reference frames [7],
2. verification of depth estimates [24],
3. proximity of a candidate target to cameras [1],
4. expected reduction of uncertainty for range information involving individual points [35] or a model fit [40],
5. unexplained image regions [11], and
6. locations which may aid in navigation through known or unknown environments [41].

In addition, most of these methods implicitly seek to reduce overall computational complexity by reducing the total number of fixations. A few of these systems have begun to incorporate information at different levels of resolution [9]-[11], [15], [32]. Table I summarizes the methods described here.

A major difference between human vision and these computational approaches is that humans are much more adept at incorporating high-level information into the target selection process. Human scanning behavior routinely depends on the fast recognition of familiar objects at different levels of resolution [9]-[11], [15], [32]. Table I summarizes the methods described here.

Active vision involves the selective gathering of information about the 3D environment. For human visual behavior, this process is characterized by two complementary capabilities: the ability to fixate or track a point of interest, and the ability to select new fixation locations for fast attentional shifts. We use the term selective fixation to refer to this latter capability. This paper has summarized relevant research in human vision, and has examined several computational methods of selective fixation. These methods involve several visual and nonvisual criteria, and the resulting sensing strategies are highly goal dependent. Research in this direction shows promise for the future development of machine vision systems which can successfully cope with difficult real-world tasks.

**References**


A. Lynn Abbott received the B.S. degree from Rutgers University in 1980, the M.S. degree from Stanford University in 1981, and the Ph.D. degree from the University of Illinois in 1990, all in electrical engineering. From 1980 to 1985 he was a Member of Technical Staff at AT&T Bell Laboratories, Holmdel, NJ. He is currently an Assistant Professor in the Bradley Department of Electrical Engineering at Virginia Polytechnic Institute and State University, Blacksburg, VA. His research interests include computer vision, robotics, artificial intelligence, and computer architecture.

**Out of Control**

Hopelessly caught in a Camel Clutch hold and staring at death right in the face, Zogan the Bulk sees the entire slide collection from the previous week’s adaptive control short course flash by in front of his eyes.

August 1992