The Role of Fixation and Visual Attention in Object Recognition: Or, Why Stereo Vision is Not Always About 3D Reconstruction

W. Eric L. Grimson & Aparna Lakshmi Ratan

Massachusetts Institute of Technology, Artificial Intelligence Laboratory

545 Technology Square, Cambridge, MA 02139

email: welg@ai.mit.edu

Abstract

A common assumption of stereo vision researchers is that the goal of stereo is to compute explicit 3D information about a scene, to support activities such as navigation, hand-eye coordination and object recognition. This paper suggests reconsidering what is required of a stereo algorithm, in light of the needs of the task that uses its output. We show that very accurate camera calibration is needed to reconstruct accurate 3D distances, which may be difficult to attain and maintain. We further argue that for tasks such as object recognition, separating object from background is of central importance. We suggest that stereo can help with this task, without explicitly computing 3D information. To demonstrate this idea, we describe an active and attentive vision system which can efficiently and reliably recognize a target object in a cluttered scene. Our system integrates visual cues like color and stereo to perform figure/ground separation, yielding candidate regions on which to focus attention. Within each image region, we use stereo to extract features that lie within a narrow disparity range about the fixation position. These selected features are then used as input to an Alignment-style recognition system. We show that visual attention and fixation significantly reduce the complexity and the false identifications in model-based recognition. We also demonstrate that stereo can be used effectively as a figure/group separator without the need for accurate camera calibration.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-91-J-4038. The author was also supported by NSF contract IRI-8900267.

1 Introduction

The subtitle of this article is, of course, deliberately provocative, in part to capture the reader's attention, but in part also to make a point. A common assumption of researchers working in stereo vision has been that a key goal of stereo is to compute explicit 3D information about a scene, in order to support activities such as navigation, hand-eye coordination and object recognition.

While there are applications in which such information can be accurately computed, these domains require very accurate camera calibration information. We suggest that in many applications, it may be difficult to attain and maintain such accurate information, and hence we suggest that it may be worthwhile to reconsider what is required of a stereo algorithm, in light of the needs of the task that uses stereo's output. In particular, we examine the role of stereo in object recognition, arguing that it may be more effective as a means of separating objects from background, than as a provider of 3D information to match with object models.

It has been previously argued (e.g. [10]) that a central component to effective object recognition is to perform figure/ground separation before model matching. This segmentation needs to produce features that are likely to come from a single object, with minimal amounts of spurious clutter. Here, we argue that effective segmentation can be achieved if several independent cues are used in conjunction. One of the advantages of using multiple cues for segmentation is that we do not need the individual cues to be very accurate. We demonstrate, for example, how simple color measures can be used for rough segmentation and how stereo can be used for figure/ground separation without explicitly computing 3D distances and without the need for accurate camera calibration. In order to integrate cues for segmentation, we use mechanisms of visual attention to focus the resources of the recognition system onto relevant data subsets (e.g. [50, 12, 48, 26]), with particular interest in the role of fixation in object recognition. Thus, our goal is to combine several cues to achieve figure/ground separation and use active vision metaphors (e.g. [4],[1],[26],[28],[5],[31]) to direct the segmentation. Many researchers in active vision have focussed on the low level mechanisms of fixation and vergence control (e.g. [39],[40],[27],[7],[28],[26],[43]). In this work, we are interested in using active vision to facilitate the higher level task of object recognition (see also [30] for a different approach to using active vision to support recognition).

To support this argument, we examine the role of fixation and visual attention in object recognition, by building and testing an active vision system which can efficiently and reliably recognize a target object in a cluttered scene. Our system integrates visual cues like color and stereo to perform figure/ground separation, yielding candidate regions on which to focus attention and recognize target objects, using an Alignment style recognition system.

2 Some Stereo Puzzles

It has been common in recent years within the computer vision community to consider the stereo vision problem as consisting of three key steps (e.g. [21, 22]): identifying points in one image, finding matching points in the other image and using the measured disparity and relative camera orientation to compute the depth. These steps are repeated for a large number of points, leading to a 3D reconstruction of the scene, at those points. It has been commonly assumed for some time that the hard part of the problem is solving for the correspondence between left and right image features. Once one knows which points match, it has been assumed that measuring the disparity is trivial, and that solving for the distance simply requires using the geometry of the cameras to invert a simple trigonometric projection.

Intriguingly, this view of reconstructing explicit depth appears to be at odds with some psychophysical stereo demonstrations, such as in [2] or [46]. These demonstrations seem to indicate that human observers are consistently fooled into making depth judgment errors, and both [2] and [46] use this to argue that depth is not computed directly in humans, but is reconstructed from non-zero second differences in depth. As a consequence, they demonstrate that human stereo vision is blind to constant gradients of depth. Similar observations on the role of disparity gradients in reconstructing depth are given by [54]. Of course, there are many possible interpretations of such psychophysical observations, and, of course, even if such observations are valid, it need not be the case that machine stereo systems make the same "mistakes" as human observers. From our perspective, we are not trying to exactly duplicate the human visual system, but we find that the existence of such illusions for humans raises an interesting question about the basic assumptions of approaches that reconstruct distance.

Consider a second puzzle about the approach of matching features, then using geometry to convert into depth. As noted, for years stereo researchers have assumed that the correspondence problem was the hard part of the task. Once correct correspondences were found, the reconstruction was a simple matter of geometry. This is true in principle, but it relies on finding the intrinsic parameters of the camera systems and the extrinsic parameters relating the orientation of the two cameras, or in otherwise solving for the camera projection models and their interrelationship. While solutions exist for finding these parameters (e.g. [51]), such solutions appear to be numerically unstable [55, 53]. If one does not perform very careful calibration of the camera platform, the result will be very noisy reconstructed distances.

Of course, there are circumstances in which careful calibration can be performed, and in these cases, extremely accurate reconstructions are possible. A good example of this is automated cartographic reconstruction from satellite imagery, where commercial systems can provide maps with accuracy on the order of a few meters, from satellite photography [18]. On the other hand, if the cameras are mounted on a mobile robot that is perturbed as it moves through the environment, then it may be more difficult to attain and maintain careful calibration. We will explore the calibration sensitivity issue in section 3, showing that small errors in calibration can have a large effect on computed depth.

Thus, we see that there are some suggestions that human observers do not reconstruct depth, and some suggestions that one needs very careful calibration (which is often hard to guarantee) in order to do this. Given this puzzle, it is worth stepping back to ask what one needs from the output of a stereo algorithm. Aside from specialized tasks such as cartography, the two standard general application areas are navigation and recognition. Interestingly, Faugeras [8] has recently argued that one can construct and maintain a representation of the scene structure around a moving robot, without a need for careful calibration. Moreover, the solution involves using relative coordinate systems to represent the scene, so that there is no Euclidean reconstruction of the scene. Robert et al.[41] have presented a navigation system that uses weakly calibrated stereo images (i.e. images for which only the epipolar geometry is known) without the need for 3D reconstruction. The theme of using projective invariants to obtain relative 3D structure with weakly calibrated cameras has also been explored in [34, 38, 36, 37, 42, 41], with primary emphasis on navigation.

What about object recognition? We have found it convenient to separate the recognition problem into three pieces [10], namely, **selection** (to extract subsets of the data features likely to have come from a single object), **indexing** (to look up those object models that could have given rise to one such selected subset) and **correspondence** (to determine if there is a way of matching model features to data features that is consistent with a legal transformation of the model into the data).

We have argued [10] that for many approaches to recognition, the first stage is the crucial one. In many cases, it reduces the expected complexity of recognition from exponential to low-order polynomial, and in many cases, it is necessary to keep the false positive rates under control. If we accept that the hard part of recognition is selection, rather than correspondence, then this has an interesting implication for stereo. If stereo were mainly oriented towards solving the correspondence problem, it is natural to expect that it needs to deliver accurate 3D data that can be compared to 3D models. But if stereo is mainly intended to help with the selection problem, then one no longer needs to extract exact 3D reconstructions, one simply needs stereo to identify data feature subsets that are roughly in the same depth range, or equivalently do not have large variations in disparity. We will examine a modified stereo algorithm in section 4 that takes advantage of this observation.

If one accepts that stereo is primarily suited for segmentation, not for 3D reconstruction, this leads to the further question of whether recognition of 3D objects can be done without explicit 3D input data. A number of recent techniques have shown interesting possibilities along these lines, for example, the linear combinations method [52] suggests that one could use stored 2D images of a model to generate an hypothesized 2d image which can then be compared to the observed image. Again, one does not need to extract exact 3D data.

We will use this to argue that stereo can contribute to the efficient solution of the object recognition problem, without the need for accurate calibration and without the need for explicit depth computation. In this case, the importance of eye movements or related control strategies is increased, causing us to reexamine the structure of stereo algorithms.

3 Why Reconstruction is Too Sensitive

While our first point is based primarily on earlier psychophysical observations, the second point bears closer examination. Let's look in more detail at the problem of computing distance from stereo disparity. In [11], we consider a detailed analysis of the sensitivity of depth reconstruction from stereo disparities to the accuracy of camera calibration. Here we note the main results.

If we are given a camera geometry with baseline b, and two cameras symmetrically verged (i.e. each camera makes an angle of γ, or -γ with respect to the perpendicular to the baseline), each has a focal length f, and the disparities are d_ℓ,d_r (i.e. the offset of the projected point in each image from the respective projective centers), then the following approximation for the computed depth is valid.

If we represent the computed depth Z at a point in terms of interocular spacing, by using ($Z' = \frac{Z}{2b}$) and use disparities as angular arcs by using ($d'_r = \frac{d_r}{f}$, $d'_\ell = \frac{d_\ell}{f}$), then, we have [11]

$$Z' \approx \frac{1 + \gamma(d'_r - d'_\ell)}{2\gamma - (d'_r - d'_\ell) + 2\gamma d'_r d'_\ell}. \tag{1}$$

Our concern is how uncertainty in measuring the camera parameters impacts the computed depth. Here, we have focused on methods that explicitly rely on individual parameters of the camera geometry. There are other methods that directly solve for the camera projection models, without explicitly extracting camera parameters. We have not had an opportunity to perform a similar stability analysis on such methods, although we expect they may also have similar difficulties.

Ideally, we would like a linear relationship between computed parameters and reconstructed depth, so that, for example, a 1% error in computing a parameter would result in at most a 1% error in depth. In [11], a perturbation analysis of this analysis shows that three key parameters can lead to significant errors in computed depth. First, the analysis in [11] shows that errors in locating the two principal points can lead to large errors, e.g. for an object 1 meter away from our standard camera setup, errors on the order of 10 pixels lead to 10% errors in computed depth. Common methods for locating principal points [19] have residual errors on the order of 6 pixels.

Errors in computing focal length lead to small errors in computed depth, for example, a 1% error in estimating f and disparities on the order of 10 pixels, will still only lead to 1% errors in relative depth for nearby objects ($Z/2b \approx 10$), which is small. Note that as the disparities get larger, the error increases. This has the interesting implication that if the object of interest is roughly fixated (i.e. the two optic axes intersect at or near the object) then disparities for features on the objects will be small, and the depth error will be small, while objects at larger disparities will have larger errors. Note that a similar observation has been made by Olson [25] who shows that much of the sensitivity of depth reconstruction to camera parameters can be isolated in the computation of the depth of the fixation point, while relative depth of other points with respect to this fixation can be computed fairly accurately.

Errors in estimating the gaze angle are more problematic. Here, a 1 degree error in estimating the gaze angle will lead to 34% relative depth errors for nearby objects ($Z/2b \approx 10$), and even a .5 degree gaze angle error will lead to 17% relative depth errors. This is graphed in more detail in Figure 1. Similarly, in Figure 1, we plot the accuracy in gaze angle needed to keep the relative depth error at most 1%, as a function of distance to the object. We note that errors due to gaze angle calibration could be a real problem. It is interesting to note that the human system appears able to measure gaze angle only up to an accuracy of roughly 1 degree [13].

In short, we need to be certain that we have estimated the principal points accurately, and that we have very accurate measurements of the gaze angles of the cameras. If we cannot do so, then we will suffer distortion in our computed depth. More importantly, that distortion varies with actual depth, so the effect is non-linear. If we are trying to recognize an object whose extent in depth is small relative to the distance to its centroid, then the effect of this noise sensitivity is reduced. This is because the effect of the error will be systematic, and in the case of small relative depth, this uncertainty basically becomes a constant scale factor on the computed depth. On the other hand, however, if the object has noticeable relative extent

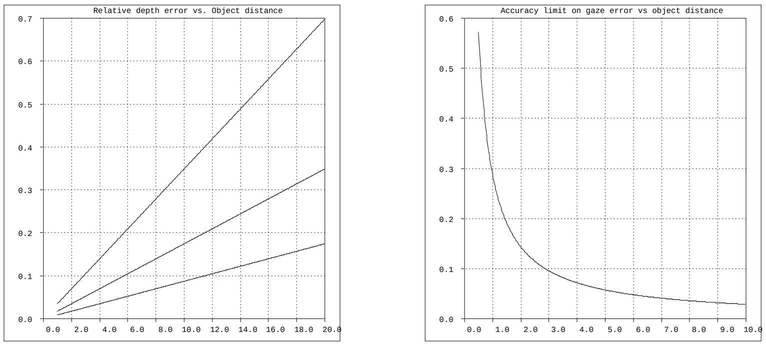

Figure 1: (left) Plots of the percentage error in computing depth, as a function of the distance to the object (in units of interocular separation). Graphs are for errors in computing the gaze angle of 1, .5 and .25 degrees, from top to bottom. (right) Plots of the accuracy in gaze angle (in degrees) needed, as a function of the distance to the object (in units of interocular separation), so that the relative error in depth is less than 1%.

in depth (even on the order a few percent), then the uncertainty in computing depth will skew the results, causing difficulties for most recognition methods that compare computed 3D structure against stored models. Thus, the sensitivity may cause serious problems for recognition methods, both due to the large errors in depth and due to the distortions with varying depth.

4 Another Look at Stereo

Given that it may be difficult to reliably compute distance, and that distance may not be needed to handle the two main uses of stereo output, we suggest that it is useful to reconsider the performance requirements that stereo should satisfy to support tasks such as object recognition. To handle figure/ground separation, a stereo algorithm should be able to do the following: detect proximal (in the image) features that lie within some range of depth [1], be able to align matching distinctive features so that they are centered in the two images, ensure that nearby parts of the corresponding object are visible in both images and can be matched, and integrate other visual cues about possible trigger features to foveate and fixate. We note that such cues are not central to the performance of a stereo algorithm. We argue, however, that since our focus in on figure/ground separation, it is important that stereo be easily integrated with other such cues, and we explore this issue here.

First, we should consider whether we can use existing stereo algorithms (e.g. [9], [29], [15])

[1]find points that are near one another in 3D space, even if one does not know exactly where in 3D

to tackle the problem of figure/ground separation. Most current stereo algorithms solve the correspondence problem as follows: Given any left image edge, search the set of right image edges for a unique match. The search is usually constrained by the (assumed known) epipolar geometry, and by a set of similarity constraints (e.g. edges should have similar orientation, similar contrast (or intensity variation), and so on). This holds both for matching individual edge points (in which case additional constraints such as figural continuity may also apply) and for extended edge fragments.

The key question is what constitutes a unique match, and this depends on the control mechanism used by the algorithm. For example, most of these algorithms attempt to find matches over a wide range of disparity, reflecting the fact that the viewed scenes may have objects ranging from close to the viewer (less than 1 meter) out to objects at the horizon. This can easily translate into disparity ranges on the order of several hundred pixels. The problem is that under these circumstances, it may be very difficult to guarantee uniqueness of match, especially when one is only considering local attributes of features, such as orientation and local contrast. One solution is to incorporate local geometric information about nearby edges [3], [24]. But an alternative is to consider changing the control mechanism.

The key problem is that previous stereo algorithms had as their goal the reconstruction of the scene, and hence they were designed to find as many correct matches as possible, over a wide range of disparities. On the other hand, if all we are interested in is separating out candidate image features that are likely to correspond to a single object, and we are willing to allow edge features to participate in several such groups, then an alternative control method is viable. In particular, since we are interested in finding roughly contiguous 3D regions, it is attractive to envision a control method in which one fixates at some target, then searches for matching features within some range of disparity about that fixation point, collecting all such matching features as a candidate object, and continues. Such an algorithm is similar in approach to some earlier stereo methods, notably [21], [22] [3], and it bears some similarity to evidence of the human stereo system, particular in the restriction of matching disparities only over a narrow range about the fixation point (referred to as Panum's limit in the perceptual literature) and the role of eye movements in guiding stereo [21, 22, 25].

To demonstrate this idea, we have implemented an active stereo algorithm (influenced in part by earlier algorithms [3, 24, 25]), which we describe in the following sections. We also indicate one method for coupling a figure/ground separation stereo algorithm with other figure/ground cues.

5 An active stereo system

Our active attentive visual system consists of a two camera, eye-head system which can pan and tilt and which lets each camera verge. The system has a pan range of $\pm 75^o$, a tilt range of -80^o to 90^o and individual eye vergence range of $\pm 25^o$. We describe the elements of the system in more detail below.

The goal of the system is to efficiently find a target object in a cluttered environment with minimal false positives. Scanning the entire field of view at high resolution is impractical, so we use a coarse-to-fine strategy whereby we use visual cues to quickly isolate regions of the image that are likely to contain the object. There are many cues that can be used for this purpose. They could be shape based cues like edges and 3D shape information from motion and stereo or they could be region based cues like color, texture, etc. Since most recognition techniques use shape based cues, we need to extract such shape based features from the scene before recognition. However, shape based cues are not sufficient to control the complexity of the problem when used on their own. Consider a case where we have $m = 50$ model features and $n = 500$ data features. If we use Alignment-style recognition techniques (e.g. [14]) where we find 3 corresponding points between the model and the image to compute the transformation that aligns the model with a hypothesized instance of the object in the image and then verify the hypothesis by comparing the transformed model with the image data, then the number of alignments that have to be tried is $O(m^3 n^3)$ which gives us on the order of 1.5x10^{13} cases to be tried in this example. Even for other types of recognitino methods, similar complexity issues will hold. Thus, we see that we need more information to control the number of alignments that have to be tried. We can use information provided by the region based cues like color and texture to reduce the number of features that have to be tried. Region based cues are useful in selection (figure/ground separation) since they provide us with a method for roughly comparing properties of the model with the data, so that we can exclude extraneous information without ignoring any relevant target regions. In this system, we investigate the possibility of combining cues to direct the fixation of the eyes on candidate regions and analyze these regions at a finer level to select out target features that can be fed into a recognition engine. Another advantage in using multiple cues is that the individual cues can be inaccurate, while still supporting reliable figure/ground separation.

In our system, we use Alignment-style ([14], [52]) recognition techniques to find targets in a large room. The room is scanned by pan and tilt movements of our eye-head system to select different viewing points. We use color information and stereo edge information to reduce the number of target features that have to be verified by the recognition system.

5.1 Color to preselect focal regions

While we could just use shape based cues like stereo and motion to achieve selection for model-based recognition, we would like to demonstrate the effectiveness of combining shape based cues like stereo with region based cues like color in controlling the combinatorics of recognition methods. Shape cues, in contrast to color, tend to be highly resolution dependent and extracting shape dependent features (e.g. corners) may require elaborate processing. Here, we describe the simple method used to extract rough regions in the image that could contain an instance of the target object based on its color properties. Our simple algorithm causes false positive identifications but we can tolerate these errors in the system since color is used in conjunction with stereo and the combination of cues helps to weed out some of these false targets.

5.1.1 Color Labeling Algorithm

The target object is modeled by building histograms of its component colors and representing each color by 6 values which correspond to the mean hue, saturation and value and the standard deviation of the hue, saturation and value from the mean. If an object has very different color distributions from different viewpoints, we should build multiple representations of that object, each described by the color distribution in that view. For the examples shown in this paper, we have simplified our task by using objects with color distrbutions that are consistent over most views of that object.

The algorithm to preselect regions in the image is a simple sequential labeling algorithm which finds the connected components in the image that match the color description of the model and represents the connected components by best fit ellipses. Since we do not attempt to model color constancy, we assume that the color of the light source does not change drastically in our experiments. While this simple algorithm has been sufficient to illustrate the importance of color in selecting regions to focus attention, we can generalize it to histogram matching approaches to color segmentation (e.g. [47]) or other color saliency algorithms (e.g [49]) to obtain the same results.

5.2 Using Color and Stereo in a Recognition System

1. An outer loop scans the area using a head eye system with small increments in pan and tilt angle.

2. At each head position, a coarsely sampled left and right image are extracted.

3. Linear edge segments are extracted from both images.

4. Each feature is a line segment and is described by its normal to the edge offset from the origin, base and end points, tangent from base to end, length of edge, mean intensity on either side, mean hue and saturation on either side, right and left neighbors, and a list of possible matches.

5. Using the color information of the target object, the left and right images are segmented into regions (ellipses) that could contain the target object, i.e. regions that have roughly the same color as the target object.

6. Keep only the features that fall within the salient regions extracted above.

7. Potential matches between features in the left and right images are computed in the following way using a stereo matcher [10] that is described in greater detail below.

   Every pair of matching features in the left and right images must have the same contrast sign (whether there is a dark to light transition or a light to dark transition at the edge), roughly the same orientation, sufficient epipolar overlap with the right edge, roughly the same intensity, hue and saturation values on at least one side of the edge, roughly the same arrangement of neighboring edges at one of the endpoints.

   We have two uses of stereo in our system. One is to find all stereo matches for features, within some depth range. The second is to find unique focal features, i.e. features with a unique stereo match over a wide range of disparity. In the latter case, in addition to the matching constraints given above, the algorithm takes advantage of the following global constraints: the continuity constraint (a good match will be supported by many other neighbouring good matches) and the uniqueness constraint which says that there can be only one correct match along the left and right lines of sight[22].

   If the focal edge in the left image has only one match in the right image, this is accepted as the correct match. Otherwise, if the left edge has more than one match in the right image, the algorithm scans a neighborhood about the ambiguous match, looks at nearby matched segments and accepts the best match based on the recorded matching information.

   Thus, the stereo algorithm finds distinctive segment-features (that lie in the salient color regions from step 5) in the left image which have unique matches in the right image, as measured over the full range of possible disparities. The distinctiveness is measured as a combination of the length and contrast of the feature. Such features could serve as focal trigger features which can be used to fixate the cameras.

8. The disparities associated with each target (trigger) edge in the left image and its matching edge in the right image are used to pan and tilt the head so that the corresponding 3D feature is centered between the cameras.

9. A finely sampled (high resolution) pair of images is taken and information around the matching features is extracted from these images. Salient regions are selected using the color properties of the model as in step 5 and edges within the salient regions are extracted. The edges are matched under the restriction that a match is sought only to within a small depth of field about the fixation position. All edges that have a match at this narrow depth of field, together with their neighboring edges (edges that lie close to them) in the image form the input to the recognition engine. If the object has significant depth variation, we can allow our system to fixate a number of locations about the target fixation, and combine the matched features over that range of depth exploration.

10. Alignment [14] is used to determine if the target object is present among the selected edges. The results of the alignment are saved and step 8 is repeated with the next trigger feature for the cameras to fixate (from step 7). Once all the trigger features have been fixated in turn, the best result alignment result is saved. If this result indicates the presence of the object in the scene, the system returns the model aligned with the image, otherwise the system returns with a message that it could not find the target object in the scene.

The system is used to find a target object starting with a given head position and going through the steps mentioned above. The cues used at present are color and stereo, although one can clearly extend the method to incorporate other cues. The final output is the model aligned with the image or a message that the target object was not found in the scene.

5.3 Experiment showing selection using color and stereo. Figure 2–Figure 4

Figure 2 shows the results of selection on a scene where there are two red objects (car and the water gun that are separated in depth but are connected in the image). The second row shows the segments and the results after running the color algorithm. We see that the color algorithm returns both objects as belonging to one connected component. The coarse level stereo finds features to focus attention and center the subsequent processing. The selected features obtained in 3 fixations are shown in figures 3 and 4. We see that in the process of fixating different features, the second stereo match yields features in a narrow depth range that come from single object. Thus, in this example stereo selects and groups features coming from single objects after the color algorithm grouped both objects in the same connected region. The selected features are verified by an alignment-style recognition engine.

6 The Recognition System

We used an alignment-based recognition system [14, 52]. These methods use a minimal set of corresponding features to produce a transformation that aligns the model with the image. These methods tend to have problems in cluttered environments. We show the advantages of using attentional selection while using alignment-methods to recognize objects in cluttered environments. The model is represented as a set of views of the object with full correspondence between the views. The features used for alignment were corner features (intersection of line segments).

7 Description of the models and test scenes

The system was used to perform the equivalent of playing "Where's Waldo", i.e. it was required to find a small target in a cluttered environment quickly by focusing its resources on regions that are most likely to contain the target object. The targets are colored objects that are represented as a set of views. The objects did not have much texture on them and were placed in a cluttered airfield scene indoors in a lab. The scene also had other distractor objects with features resembling the model features. The distractor objects had similar color and shape as the model objects. The scene was imaged using two color CCD cameras mounted on a head-eye system. The orientation of the target and the intensity of the lights in the room were varied while testing the system. The starting position of the head was also varied so that the target object was either totally visible in both the left and right images, partially visible in both images, visible in one eye and not visible in the other eye or not visible in both eyes. Figure 5 shows a typical scene containing the model.

8 Experiments

The system was tested by running it a number of times on scenes containing different target objects, in several orientations under varying lighting conditions. Table 1 gives the average reduction in search at the various stages after 50 trials.

Experiment 1 - Figure 5

Figure 5 shows the results at different stages of one run of the system where the head was initially positioned so that the target object was present in both the left and right images The head initially takes a pair of coarse resolution images. The images are processed to find edges and approximate the edges by line segments. The color filter is applied and only those line segments that fall in regions that have color properties similar to the color properties of the target object are retained. The results of applying the simple color filter are not perfect but they are useful in guiding the stereo match by reducing the number of correspondences that have to be tried by the stereo algorithm. The stereo matching is done on the segments retained after applying the color filter and the head is turned so as to center the focal edge resulting from the stereo match in the left image. The head takes two more images and extracts the 256 * 256 array from the center of the images to obtain foveated images. The segments are extracted from these images (shown in the fourth row) and the stereo match is done in a narrow disparity range about the disparity of the focal match (fixation disparity). The resulting set of segments which are likely to contain the target object are shown in the last row of Figure 5. These selected segments are fed into an alignment-based recognizer and the results of the alignment are shown in the last row as well. The good resulting alignment indicated that the selected set of segments are an instance of the target object.

Experiment 2 - Figure 6

Figure 6 show the results at different stages of a run of the system where the head was initially positioned so that the target object was present in the image taken by the left eye and absent in the image taken by the right eye. The stereo matching is done on the segments retained after applying the color filter and the head is turned so as to center the focal edge resulting from the stereo match in the left image. The selected segments (shown in the last row in Figure 6 are fed into an alignment-based recognizer and the results of the alignment indicated that the selected set of segments did not belong to the target object.

Experiment 3- Figures 7–8

Figures 7–8 illustrate the processing at the various stages for a different orientation of the same object (plane) and different intensity of the light source. In this example, the target object is found in two fixations. Figure 7 shows the initial scene and segments before and after the color filter. Figure 8 shows the results of the first fixation when the eyes foveate a distractor object (plane) that has the same color properties as the target object. The results of the alignment which indicates that the object is not found in the selected set of segments. The head turns and fixates on the next focal edge which foveates the correct target object as seen in in right two columns of Figure 8. The target object is recognized with this fixation.

Experiment 4 - Figures 9–10

Figures 9–10 show the various stages of the processing on a different object (red car). Figure

9 shows the initial images before and after the color processing. Figure 10 shows the results of the first fixation where the head fixates and rejects the distractor object which has similar color to the target object and the second fixation where object (car) is found by the system.

9 Conclusions

We have suggested that stereo may play a central role in object recognition, but not in the manner usually assumed in the literature. We have suggested that stereo may be most useful in supporting figure/ground separation, and that to do so it need not compute explicit 3D information. Supporting this argument was the observation that depth reconstruction is extremely sensitive to accuracy in the measured camera parameters.

We demonstrated this idea using an attentive vision system that searches for objects in a cluttered scene. The system incorporates both color and stereo as cues for separating out likely regions in the scene, which are then verified using an Alignment style recognition system. The method illustrates one mechanism for accomplishing figure/ground separation and shows how the complexity of recognition can be reduced by using such a mechanism.

Using the idea of depth detectors tuned to a narrow range about a fixation point has been previously explored in the literature, primarily for obstacle avoidance [16], [27]. This work considers the same general idea within the context of recognition. Such an approach opens up several other avenues for investigation. For example, what is the role of other visual cues in aiding the stereo matching problem. While one option is to augment image features with attributes, such as texture or color measures, an alternative is to consider using such cues to drive vergence eye movements, helping to align the cameras on trigger features, so that the local matcher can extract image features likely to correspond to a single object. We intend to explore these and related issues in the near future.

Stage	Left Image	Right Image	Number of Alignments ($\times 10^8$)
Initial segments	480	450	10000
After color	108	90	100
Focal segments	500	400	10000
Selected segments	25	25	1

Table 1: This table gives the average number of segments in the left and right image together with the avg. no. of alignments that have to be tried at the various stages of the system. The average number of fixations before the target object was found is 3.

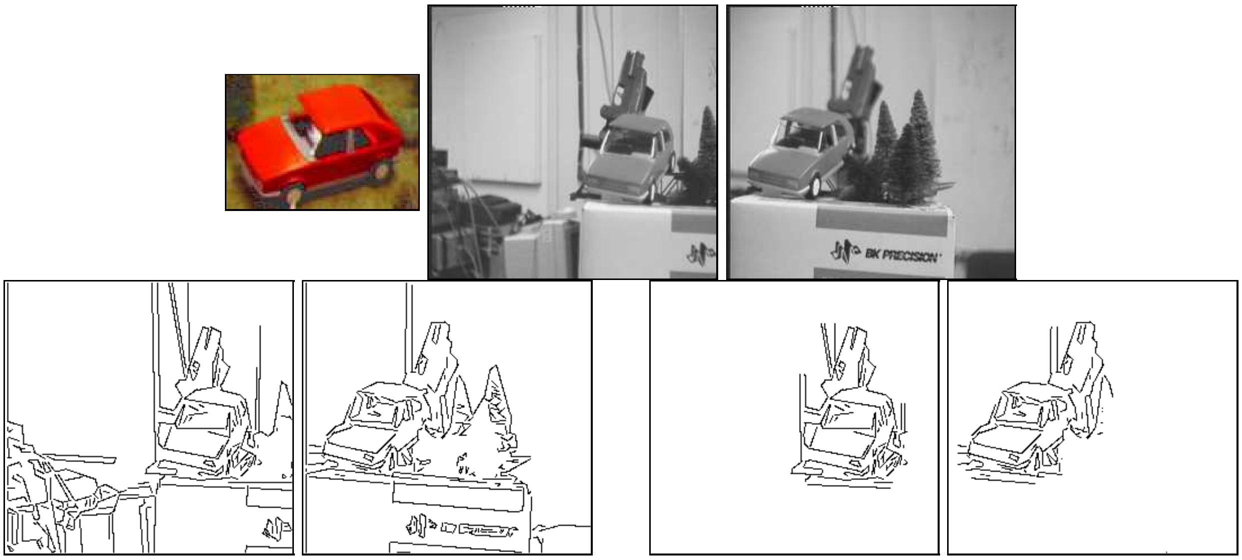

Figure 2: Top: target object, initial images. Bottom: segments before and after color filtering

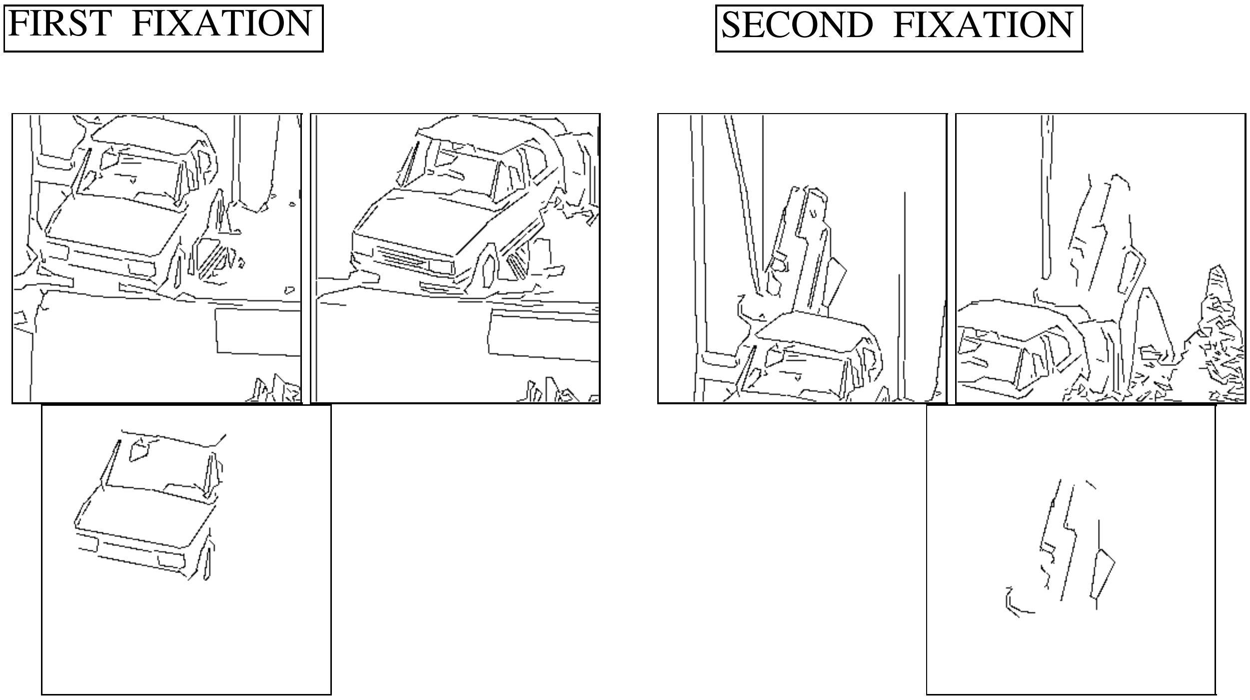

Figure 3: Selected segments in first 2 fixations.

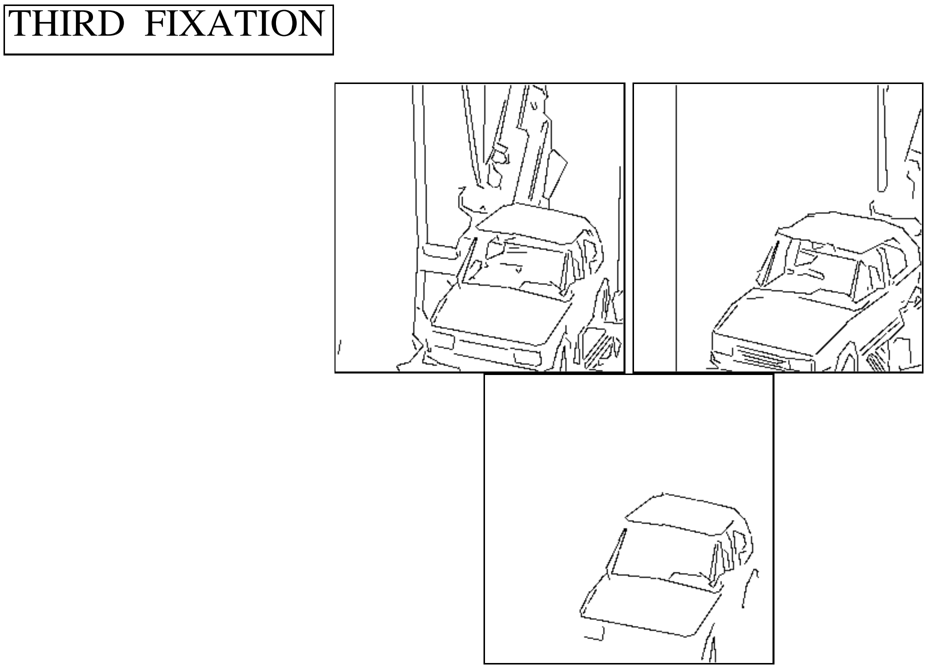

Figure 4: Selected segments in third fixation.

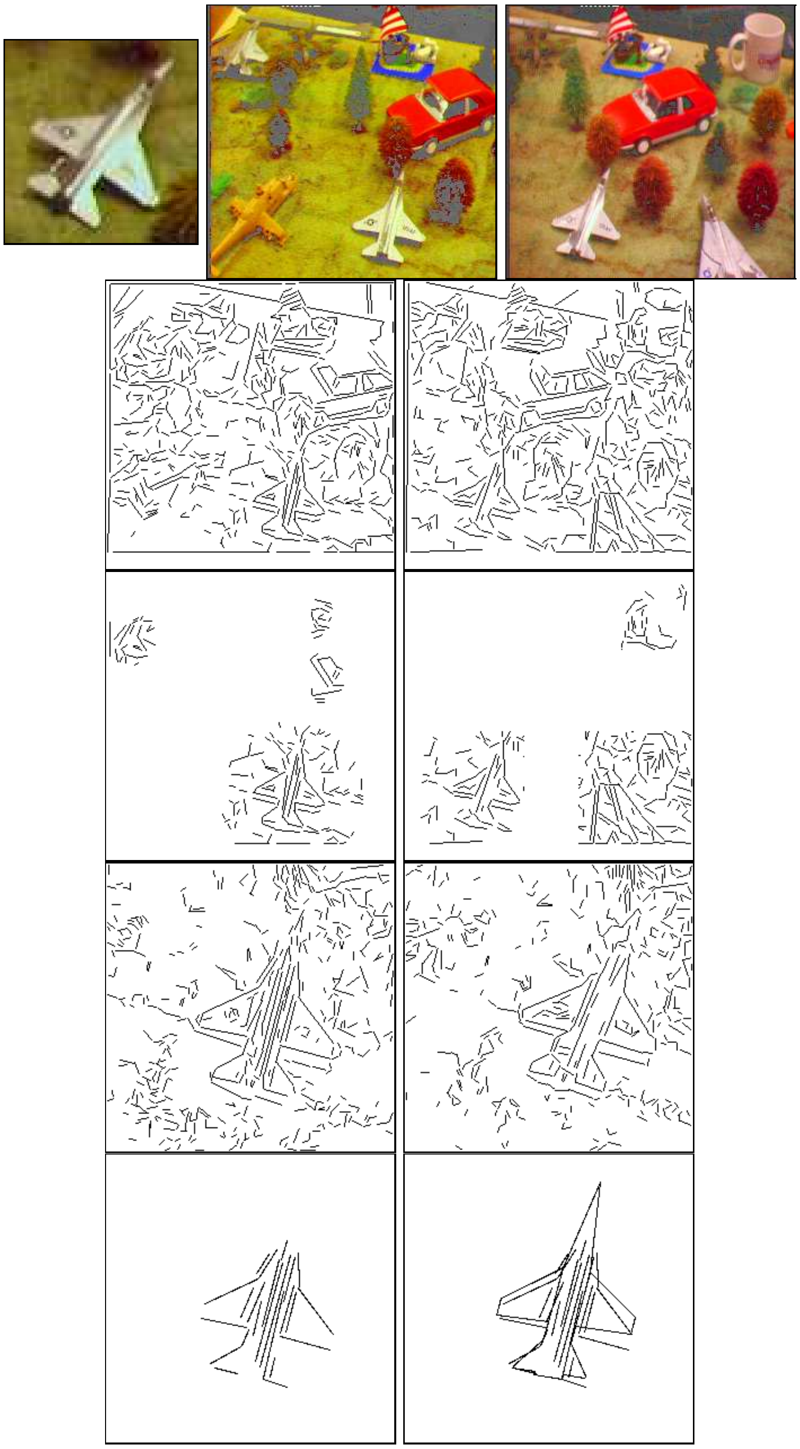

Figure 5: First two rows have target,initial images, segments. Third row has the segments after color filtering. Fourth row has the segments in foveated images and fifth row has the selected segments and the results of alignment indicating that the object was FOUND

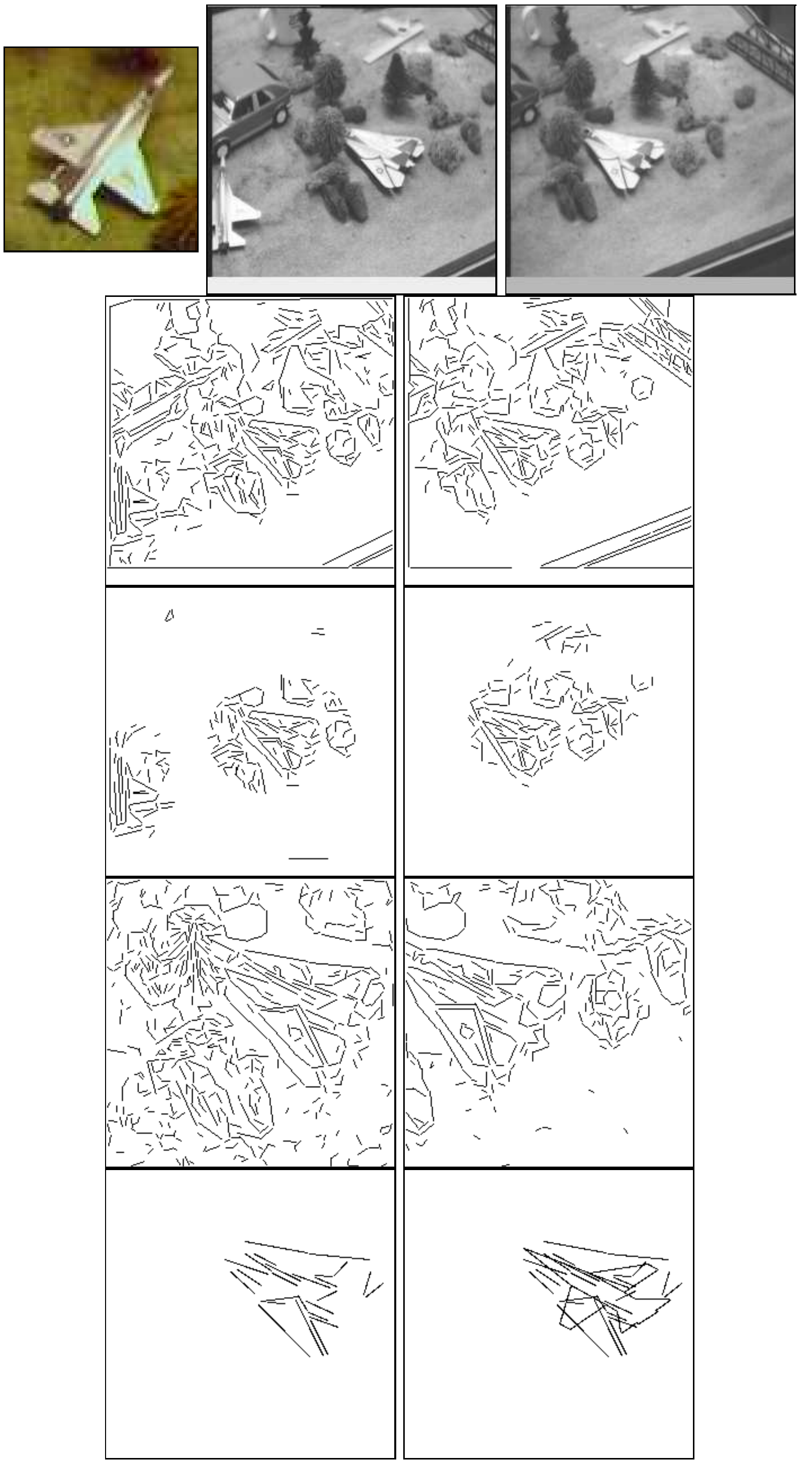

Figure 6: First two rows have target,initial images, segments. Third row has the segments after color filtering. Fourth row has the segments in foveated images and fifth row has the selected segments and the results of alignment indicating that the object was NOT FOUND

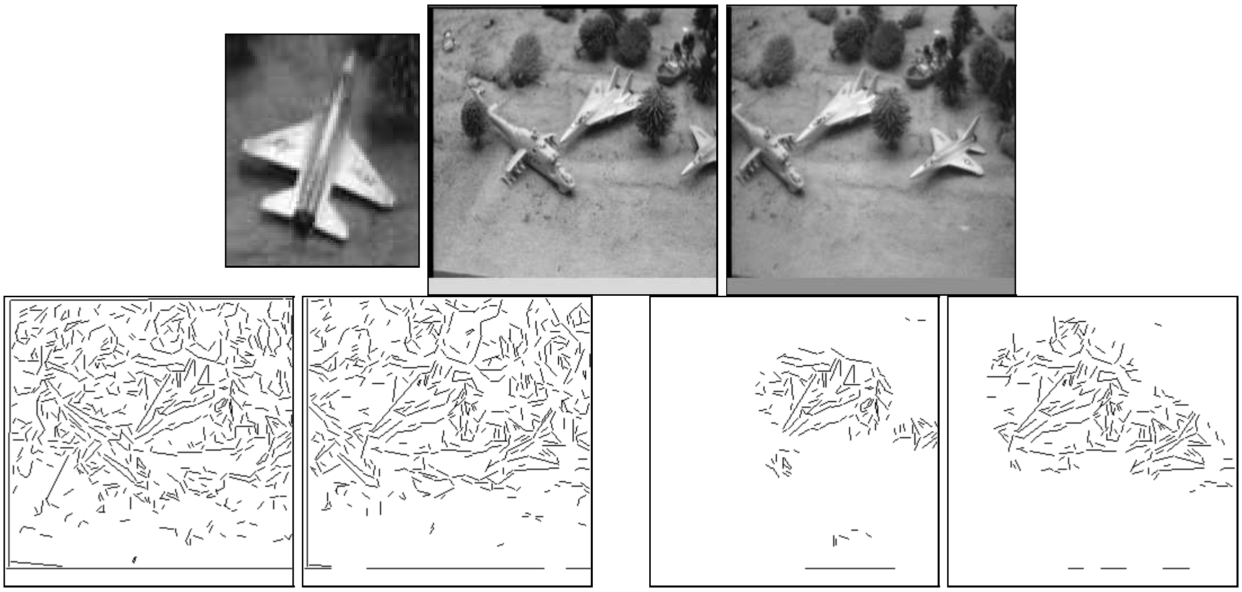

Figure 7: Top: target object, initial images. Bottom: segments before and after color filtering

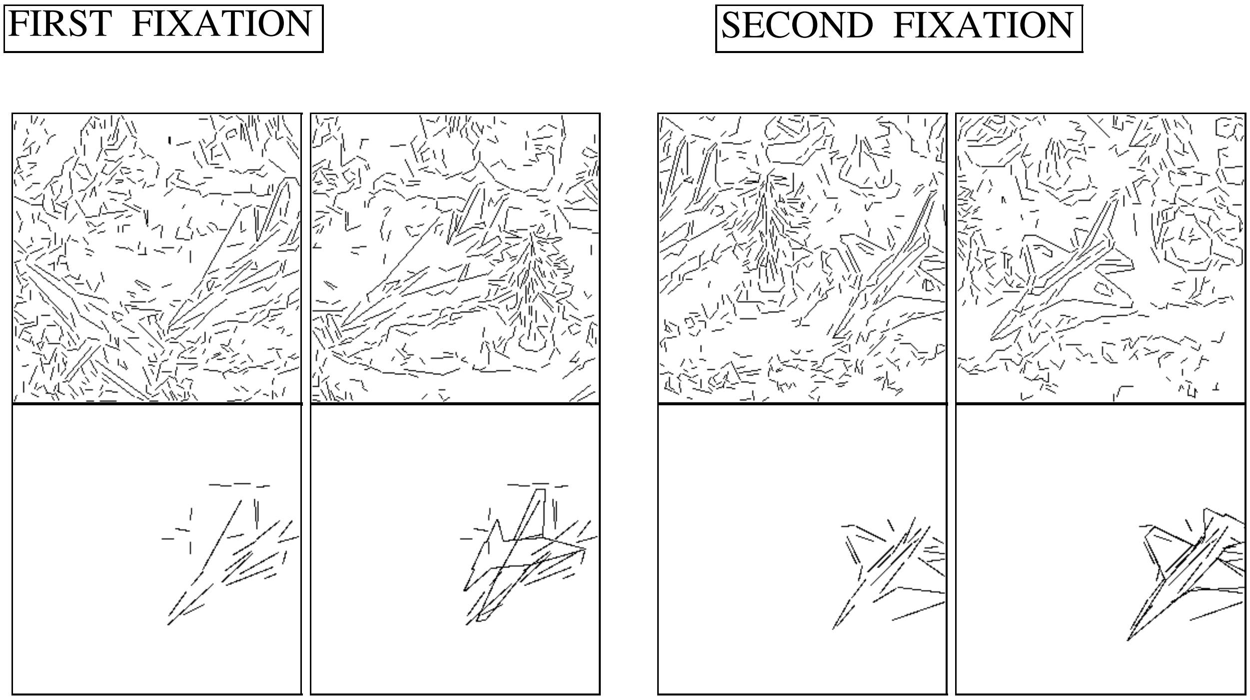

Figure 8: Selected features in two fixations. Target is FOUND in second fixation.

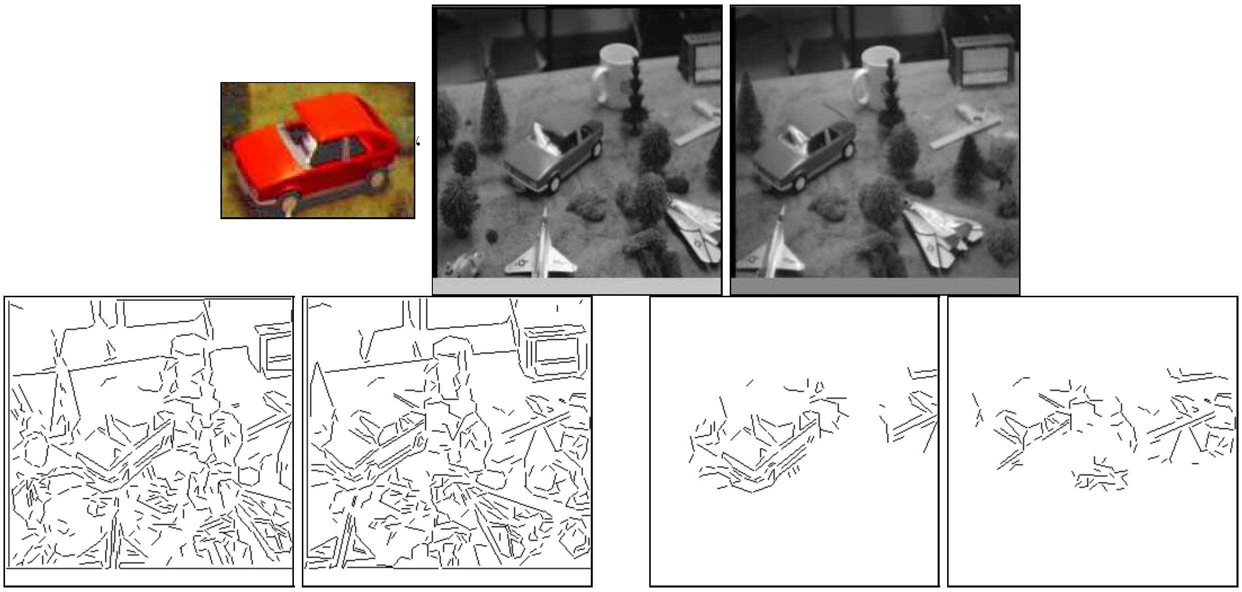

Figure 9: Top: target object, initial images. Bottom: segments before and after color filtering

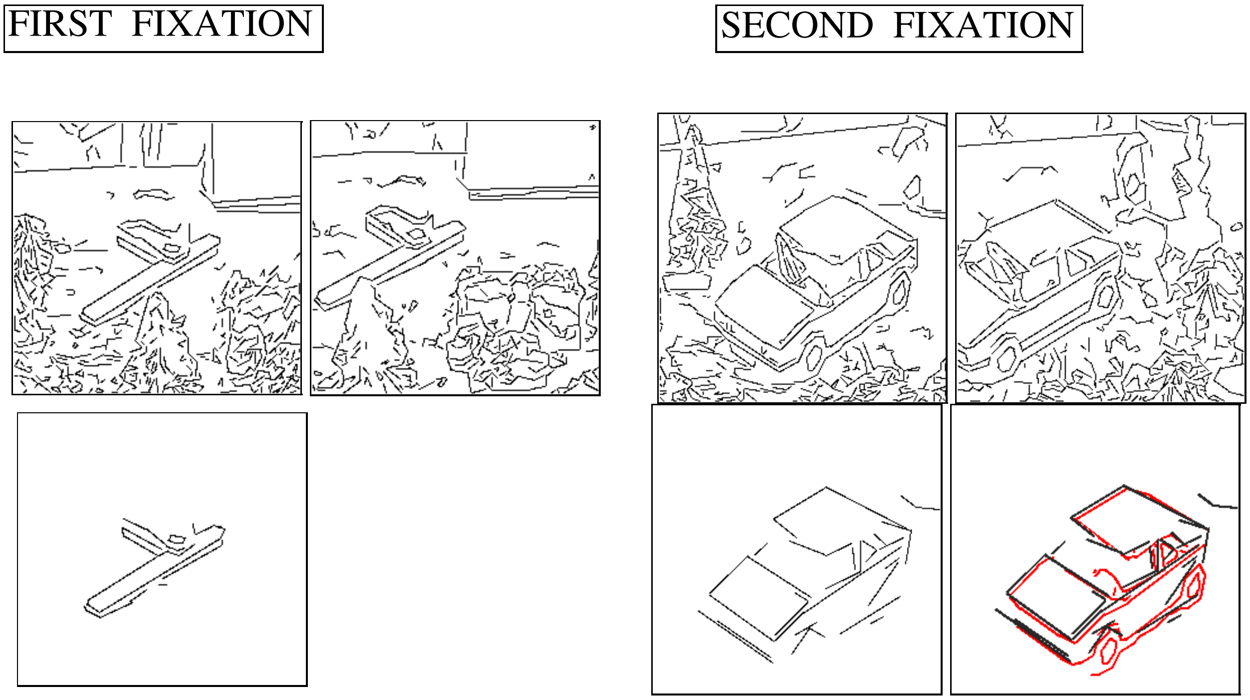

Figure 10: Selected segments in 2 fixations. Target is FOUND in second fixation.

References

[1] Y. Aloimonos, "Active Perception", *Lawrence Erlbaum Assoc., Publishers*, 1993.

[2] S.M. Anstis, I.P. Howard, and B.J. Rogers, "A Craik-O'Brien-Cornsweet Illusion for Visual Depth", *Vision Research*, **18**:213–217, 1978.

[3] N. Ayache and B. Faverjon, "Efficient registration of stereo images by matching graph descriptions of edge segments", *International Journal of Computer Vision* **1**(2):107–131, 1987.

[4] D. Ballard, "Animate Vision." *Artificial Intelligence*, Vol 48, 57-86, 1991.

[5] R.C. Bolle, "Data and model driven foveation" *ICPR*, Atlantic City, 1990.

[6] M. Bober, P.Hoad, J.Mataas, P. Remagnino, J.Kittler, J. Illingworth, "Control of perception in an Active Vision System: Sensing and Interpretation", *IROS*, 1993.

[7] D. Coombs and C. Brown "Real-Time Binocular Smooth Pursuit",*IJCV*, 11:2, 147–164, 1993.

[8] O. D. Faugeras, "What can be seen in three dimensions with an uncalibrated stereo rig?", *Second European Conference on Computer Vision*, Santa Margherita Ligure, Italy, pp. 563–578, 1992.

[9] W.E.L. Grimson, "Computational experiments with a feature based stereo algorithm", *IEEE Pattern Analysis and Machine Intelligence*, **7**, No. 1, 17–34, 1985.

[10] W. E. L. Grimson, *Object Recognition by Computer*, MIT Press, 1991.

[11] E. Grimson, *Why Stereo Vision is not always about 3D Reconstruction*, MIT AI Lab Memo 1435, 1993.

[12] C. Koch and S. Ullman, "Selecting one among the many: A simple network implementing shifts in selective visual attention", *Technical Report, M.I.T., AI-memo-770*, Jan 1984.

[13] I.P. Howard and W.B. Templeton, "Human Spatial Orientation", John Wiley & Sons, 1966.

[14] D.Huttenlocher, S.Ullman, "Object Recognition Using Alignment", *ICCV*, 109-111, 1987.

[15] W. Hoff and N. Ahuja, "Surfaces from stereo: Integrating feature matching, disparity estimation, and contour detection", *IEEE Trans. Patt. Anal. & Mach. Intell.* **11**(2): 121–136, 1989.

[16] I.D. Horswill, "Proximity Detection using a filter tuned in three-space", *DARPA Image Understanding Workshop*, pp. 973–978, 1992.

[17] D.P. Huttenlocher and S. Ullman, 1990, "Recognizing Solid Objects by Alignment with an Image," *Inter. Journ. Comp. Vision* **5**(2):195–212.

[18] Commercial brochure, ISTAR, Sophia Antipolis, France.

[19] R.K. Lenz and R.Y. Tsai, "Techniques for Calibration of the Scale Factor and Image Center for High Accuracy 3-D Machine Vision Metrology", *PAMI*, **10**, No. 5, 713–720, 1988.

[20] S.T.F. Mahmood, "Attentional Selection in Object Recognition", MIT AI Technical Report 1420, 1992.

[21] D. Marr, *Vision*, W. H. Freeman and Company, San Francisco, 1982.

[22] D. Marr and T. Poggio, "A theory of human stereo vision", *Proc. Royal Society of London B* **204**:301–328.

[23] E.S. McVey and J.W. Lee, "Some accuracy and resolution aspects of computer vision distance measurements", *PAMI* **4**(6):646–649, 1982.

[24] G. Medioni and R. Nevatia, "Segment-based stereo matching", *CVGIP*, **31**: 2–18, 1985.

[25] T.J. Olson, "Stereopsis for Verging Systems", *IEEE Comp. Vis. Patt. Recog. Conf.*, 55–60, 1993.

[26] J. J. Clark and N. Ferrier, "Modal control of an attentive vision system", *ICCV*, 524 - 523, 1988.

[27] T.J. Olson and D.J. Coombs, "Real-time vergence control for binocular robots", *DARPA Image Understanding Workshop*, pp. 881-888, 1990.

[28] K.Pahlavan, T.Uhlin, J.O.Eklundh, "Dyanamic Fixation", *Fourth ICCV* 412 - 419, 1993.

[29] S.B. Pollard, J.E.W. Mayhew, and J.P. Frisby, "PMF: A stereo correspondence algorithm using a disparity gradient limit", *Perception* **14**, 449–470, 1985.

[30] R. Rao, D. Ballard "Object Indexing using an Iconic Sparse Distributed Memory", *ICCV*, 24–31, 1995.

[31] R. Rimey and C. Brown "Control of Selective Perception Using bayes Nets and Decision Theory", *IJCV*, 12:2/3, 173–207, 1994.

[32] L. Wixson and D. Ballard "Using intermediate objects to improve the efficiency of visual search", *IJCV*, 12:2/3, 209–230, 1994.

[33] B.J. Rogers and M.E. Graham, "Anisotropies in the perception of three-dimensional surfaces", *Science* **221**, 1409–1411, 1983.

[34] A. Shashua, "Projective Structure from two Uncalibrated Images: Structure from Motion and Recognition", MIT AI Lab Memo 1363, 1992.

[35] B. Boufama, R. Mohr, and F. Veillon "Euclidean Constraints for Unclibrated Reconstruction",*ICCV*,1993.

[36] R. Mohr, L. Morin and E. Grosso "Relative positioning with uncalibrated cameras", *Geometric Invariance in Computer Vision*, 440–460, MIT Press, 1992.

[37] R. Mohr, F.Veillon and L.Quan "Relative 3D reconstruction using Multiple Uncalibrated Images", *CVPR*, 543–548, MIT Press, 1993.

[38] R.Hartley and R.Gupta "Stereo from uncalibrated cameras", *Proc. of the Conference on Computer Vision and Pattern Recognition, Urbana Champaign, Illinois*,761–764, 1992.

[39] E.P.Krotkov "Active Computer Vision by Cooperative Focus and Stereo", *Springer Verlag*, 1988.

[40] N.Ahuja and A. L.Abbott "Active Stereo:Integrating Disparity, Vergence, Focus, Aperture and Calibration for Surface Estimation", *PAMI*, 15:10, 1007–1029, 1993.

[41] L.Robert, M.Buffa and M. Hebert "Weakly Calibrated Stereo Perception for Rover Navigation", *ICCV*, 46–51, 1995.

[42] C. Rothwell, G.Csurka, O.Faugeras "A Comparison of Projective Reconstruction Methods for Pairs of Views", *ICCV*, 932–937, 1995.

[43] L. Abbott and B. Zheng "Active Fixation using Attentional Shifts, Affine Resampling and Multiresolution Search", *ICCV*, 1995.

[44] S.I.Olsen "Epipolar Line Estimation", *ECCV*, Santa Margherita Ligure, Italy, 307–311,1992.

[45] O.D.Faugeras, Q.T.Luong and S.J.Maybank "Camera Self Calibration: Theory and Experiments", *ECCV*, Santa Margherita Ligure, Italy, 321–334, 1992.

[46] K.A. Stevens and A. Brookes, "Depth reconstruction in stereopsis", *First Int. Conf. Comp. Vis.*, 682–686, 1987.

[47] M.J. Swain and D. H. Ballard, Indexing via Color Histograms. *ICCV*, 623 - 630, 1990.

[48] T.F. Syeda Mahmood, "Attentional Selection in Object Recognition", *AI Tech Report*, 1992.

[49] T.F. Syeda Mahmood, Data and Model driven Selection using color regions. *ECCV*, 321-327, 1992.

[50] A. Treisman, Selective Attention in Man. *Brit. Med. Bulletin.*, 20:12-16, 1964.

[51] R.Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses", *Journal of Robotics and Automation* **3**(4):323–344, 1987.

[52] S. Ullman,R. Basri, "Recognition by Linear Combinations of Models", *PAMI*, **13**(10):992–1006, 1991.

[53] A. Verri and V. Torre, "Absolute depth estimate in stereopsis", *Journal of the Optical Society of America A* **3**(3):297–299, 1986.

[54] R.P. Wildes, "Direct recovery of three-dimensional scene geometry from binocular stereo disparity", *PAMI*, **13**(8): 721-735, 1991.

[55] L.B. Wolff, "Accurate Measurement of Orientation from Stereo using Line Correspondence", *IEEE Conf. on Computer Vision and Pattern Recognition*, 410–415, 1989.