Using Multiple Cues for Controlling an Agile Camera Head

Claus Siggaard Andersen* and Henrik Iskov Christensen†
Laboratory of Image Analysis, Aalborg University
9220 Aalborg East, Denmark

Abstract

The interest in camera heads has been ever increasing over the past years. This interest in agile sensor systems seems promising for guiding the vision research in new useful directions. At Aalborg University, we have implemented the second and much improved version of a camera head, which specifically has lead to research in how to control the gaze mechanism. We propose a combination of multiple visual cues for controlling the camera head, enabling it to cope with less restricted scenarios than otherwise possible. By integrating accommodation and disparity cues we are able to control the vergence of the camera head separate from the version mechanism, enabling reliable smooth pursuit and gaze shift.

1 Introduction

Since the late sixties much research have been concentrated on extraction of robust and accurate 3D information.

Despite the intensive efforts put into extraction of 3D information from visual input, no one has so far been able to come up with an algorithm, acquiring the required 3D information for tasks like recognition of generic 3D objects.

Bülthoff and Mallot [3] published some experiments supporting that the human visual system integrates both monocular and binocular depth cues in order to derive a description of the environment. This includes cues such as accommodation, edge- and intensity based stereo, and shape from shading. They conclude that 3D information can be obtained by applying each cue individually, but it is the cooperation that makes the system perform so well.

This idea has also been explored by Abbott, Das and Ahuja [1, 6], who investigated the integrated use of vergence, stereo and focus. The results were inspiring, though not immediately suitable for dynamic environments.

Krotkov [7] also worked on these aspects, but did also considered the feasibility of moving the head to obtain additional, disambiguating information.

Many of the hardware related problems were overcome in the design of the head build and used at University of Rochester by Chris Brown, Dana Ballard and many others. This camera head was equipped with fast individual vergence motors and a tilt motor. Combining the head with fast image processing hardware Olson, Coombs and Brown [8, 5] designed and implemented a system capable of fixating on and tracking a moving target, in real time.

The Royal Institute of Technology, KTH, Sweden, designed a highly versatile agile binocular camera head, which currently ranges among the very best. Pahlavan et. al. [9], combined accommodation and disparity cues for controlling the vergence mechanism of the head, while the version mechanism was controlled independently. Experiments and results for both static and dynamic scenes, have been reported.

This paper addresses problems similar to the ones demonstrated at KTH, though we are taking a different approach to the extraction of cues, and the data fusion strategies differs as well. The 3D extraction process will be described in detail in the following section, along with some definitions of terms. This is followed by a section with some practical experiments and results. Finally a brief summary is given.

2 Version and Vergence

A camera head as the one used at Aalborg University, has 10 degrees of freedom, where 4 of them are rotational ones, as shown in figure 1. The rotational freedoms are referred to as tilt, pan, and gaze. We will in the following only be concerned with the "gaze mechanism". The gaze mechanism consist of two motors, enabling the cameras to be rotated inde-

*Funded by the Danish Technical Research Council under the MOBS framework
†Funded by ESPRIT Basic Research Action BR 3038/7108

0-8186-5805-3/94 $3.00 © 1994 IEEE

pendently. We will henceforth term the direction of gaze the *version* angle, and the angle between the two cameras line of sight, the *vergence* angle.

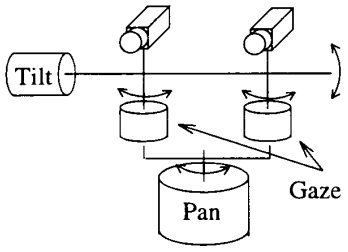

Figure 1: *An outline of the AUC camera head. The "neck" enable pan and tilt functions. The cameras themselves have independent rotation around the center of projection. In addition the cameras are equipped with motorized lenses, each with controllable focus, zoom and aperture.*

Basically two types of gaze control are required for a camera head, gaze holding (stabilization), and gaze shift. The gaze holding is the process of keeping the image of a moving object stabilized on the "retina", (i.e. center of the image plane), whereas the gaze shifting operation is the act of shifting attention from one object to another. The two different gaze mechanisms poses different requirements on the head control. Gaze shifts are performed as separate smooth vergence movements, mediated by saccadic version movements, while gaze holding requires smooth adjustments in both the version and vergence angle.

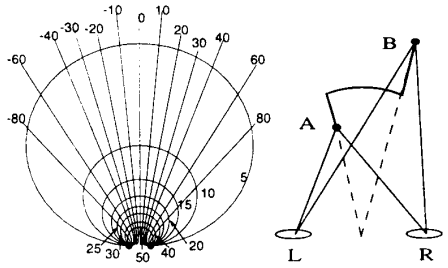

Figure 2: **Left:** *Vieth-Müller diagram. The circles correspond to positions with constant vergence. The hyperbolas corresponds to positions of constant version. The numbers are the angles in degrees.* **Right:** *The thick line represents the motion pattern of the fixation point during gaze shift from A to B.*

Carpenter [4, Chapter 5] proposes a simple and yet elegant control model for version and vergence angles in primates performing gaze shifts. This follows the structure of the Vieth-Müller diagram, shown in figure 2. Assuming equal eye dominance, the version movement will actually create symmetric projection displacements in the two eyes, as illustrated in figure 3. This will greatly simplify the following vergence movement since direction of the vergence (divergent or convergent), can be derived directly from where the projections are positioned. In case of occlusion of one eye, the mutual dominance approach has the advantage that the other eye simply becomes totally dominant.

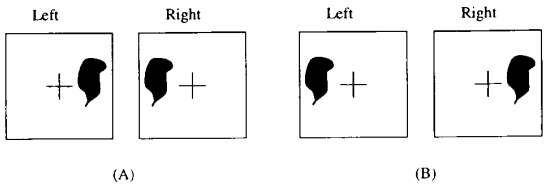

Figure 3: *The figure illustrates the symmetric vergence situation. In case (A) the the vergence angle is too large, i.e., fixation nearer the camera head is required. In case (B) the vergence angle is too small, i.e., fixation further away is required.*

However, there will be no disparity information to drive the control of the head. since only retinal slip from the non-occluded camera is available. By combining the binocular depth cue with a monocular one, the head can maintain fixation on the moving target when a single camera is occluded.

Several monocular depth cues exist, where one of the most well known is accommodation, which we will apply here.

The extraction and combination of the cues will be described briefly in the following section, followed by a section about the experimental results.

2.1 Extraction of disparity information

Since the very first algorithms for extraction of 3D information were developed, disparity based techniques have played a central role. Many stereopsis algorithms exist, but many are not applicable in a dynamic environment, where time is a critical factor. Since the disparity is to be computed in a narrow neighborhood around the optical axis, resembling foveal processing, the range of disparities is small, thus simple techniques for correspondence analysis can be used with success. Basically there exist two rivalling methods, for doing fast disparity detection in this very

restricted domain. The methods are *cepstral filtering*[10] and *phase correlation* [5]. The two techniques perform in a similar manner, and have almost identical performance, see e.g., Olson and Coombs [8] for a comparison.

Here it has been chosen to use the cepstral filtering approach, which originates from the signal processing domain, where it has been used for detecting echo signals.

The filter is based upon the notion that the left and right images, in theory at least, are horizontally shifted versions of each other, hence the relation to echo signals.

The cepstral filter is computed as the Fourier transform of the log of the Power Spectrum of the combined left and right image (spliced image) denoted by $f(x, y)$. The Cepstrum is thus given by:

$$\mathcal{C}(x', y') = \mathcal{F}(\log(|\mathcal{F}(u, v)|)) \quad (1)$$

where $\mathcal{F}(u, v)$ is the Fourier transform of $f(x, y)$. By analyzing the expression it can readily be shown that peaks will occur in the Cepstrum, at positions identifying the dominant disparities. In case of multiple disparities, it will often suffice to select the disparity corresponding to highest peak value, since the object of interest is the dominant one in the retinal visual field.

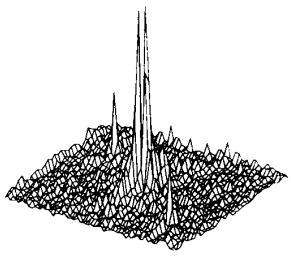

Figure 4: *A typical Cepstrum as it is computed for determining the disparity. The two peaks signify the disparity found between the images. The center peak - the DC-component has been truncated for display purposes.*

2.2 Accommodation Information

Accommodation or "depth from focus" as it is often termed, is actually two categories of techniques, that uses information about the local blur in the images to decide the proper focus distance. The first measures the degree of blur in two or more images, and from a model of the image degradation as a function of the controlled parameter an estimate of the distance to the object is obtained. The other technique is based on iteratively changing the focus distance while observing the change in the output of some specialized criterion function measuring the sharpness of the image. This function is typically designed to have a single peak for proper focus distance of a given object, which simplifies the process of finding optimal focus distance.

The first of these methods seems immediately to be the one to prefer. However, the method is usually less reliable, since it relies on a good estimate of the true blur in the image, as well as the intrinsic camera parameters.

The other method, which has been used here, has the advantage, that due to its iterative nature it is well suited for keeping track of focus during smooth pursuit.

Krotkov [7], has examined a wide variety of criterion functions, and found that a summed gradient magnitude function, known as the Tennengrad operator, performs best. On the other hand, we have already computed the FFT for the Cepstral filtering, which contains information about the image sharpness (and of course much more), and as such is an excellent mean for deciding optimal focus position.

Using knowledge about current focus setting and the current and previous criterion function values, the direction of change can be determined, and also a quantitative measure of the size of the change. An alternative is to iterate at high speed, thus the method updates the focus position fast enough to keep up with any change in targets position, although this is not pursued here.

As for the disparity cue we have chosen to run a quantitative control approach, which involves some pre-calibration of the camera head to determine lens parameters etc.

2.3 Combining Disparity and Accommodation Cues.

As it can be seen from the previous two sections, the disparity and accommodation driven vergence each have their advantages and weaknesses, which fortunately are complementary in a sense that makes the integration of them a valid suggestion.

Carpenter [4] suggests a schema for integrating the methods by fusing data according to the diagram listed in figure 5. The accommodation cue thus along with the disparity cue enables controlling the vergence mechanism and the focus position.

The actual design of the fusion process can be done in multiple ways, and still fulfill the structure in figure 5.

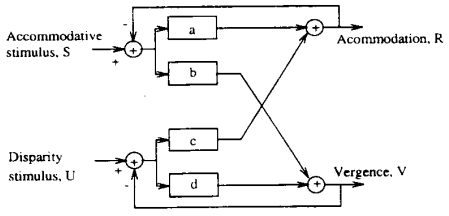

Figure 5: *A schematic layout of the control strategy proposed by Carpenter. The disparity and accommodation cues are strongly coupled into controlling the vergence and focus position.[4]*

We have decided on using a weighted averaging of the two cues. The weights are determined from a statistical model of the certainty of the two methods. Thus a combined model and data driven fusion process is derived. The statistical models was derived from an analysis of the algorithms employed, as described in [2].

3 Experiments

Two types of 3D tracking have been implemented, bright spot tracking, and template based tracking, where the former operates in a servo loop at video rate. The template based technique have been employed during the vergence experiments reported below.

3.1 Verging using disparity information

Figure 7 shows the process of controlling the vergence angle using disparity information. The disparity is converted to true distance, and feed through a PD controller, that is minimizing the vergence error. The gain factor in the PD-controller is kept artificially low in order to display the process.

Figure 7 shows the process of controlling the vergence angle using disparity information. The disparity is converted to true distance, and feed through a PD controller, that is minimizing the vergence error. The gain factor in the PD-controller is kept artificially low in order to display the process.

In figure 6 the vergence and version angles is plotted as the head is tracking a moving object.

Since the disparity signal typically is known with about one pixel accuracy, the error in the vergence

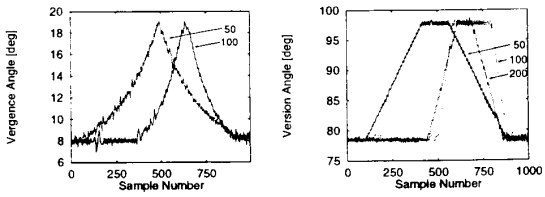

Figure 6: **Left:** *vergence angle as a function of the sample number. Object has been moved directly towards (and away from) the head. Range is approximately 1.5–3.0m, with two different velocities, 50mm and 100mm per second.* **Right:** *version angle as function of sample number. Tracking object moving parallel to the head. Horizontal identifies no movement. Velocity of 50,100 and 200 mm per second, respectively. Distance is approximately 3m.*

estimate will be proportional to the "size" of a pixel in the 3D Cartesian space, which again depends on intrinsic camera parameters and distance to the object being perceived. The certainty may be improved by using interpolation in the peak detection, as suggested by others [8].

3.2 Verging using accommodation information

The accommodation cue is based upon the high frequency content in the FFT spectrum, and the changes in this as the focus position is changed. Generally focus is a simple cues to obtain, however, the certainty of the cue is limited. Actually the certainty degrades with the square of the distance to the object, and thus the method is only valid within a rather limited range. As described earlier we have implemented the accommodation driven vergence mechanism based on a quantitative reconstruction. Thus the focus position is computed in meters, rather than just deciding if the object is closer than the current focus setting etc. This makes the method rely heavily on a good calibration of the intrinsic camera parameters, which may severely degrade the performance.

3.3 Verging by combining cues

Due to the low certainty of the accommodation cue, the accuracy of the combined approach has not been improved compared to the result for the disparity based technique. However, the method is capable of handling temporary occlusion of one camera, by transferring total dominance to the other, and then use the

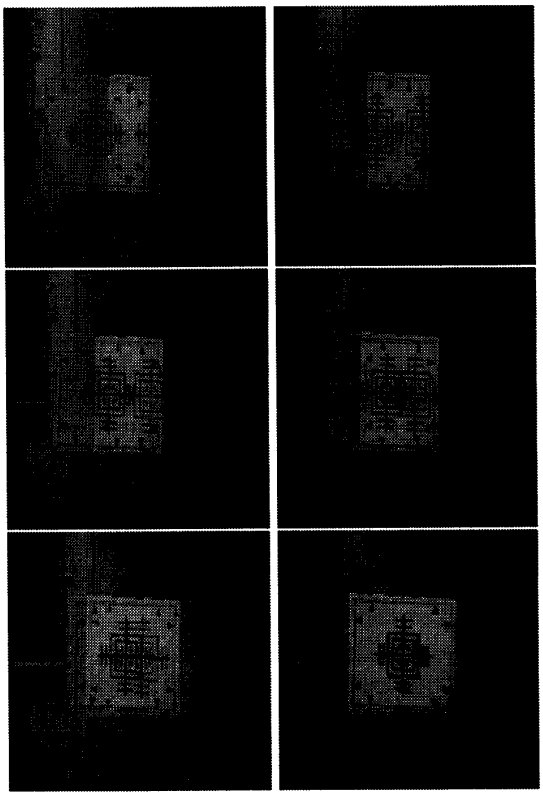

Figure 7: *Six images showing the vergence control in action. The left and right images has been overlayed to show the symmetric vergence, being issued. The image are for vergence at distances of 3.5m, 2.8 m, 2.6m, 2.4m, 2.2m, and 2.0m. The correct vergence position is 1.98 m.*

accommodation cue from that camera to control the vergence angle, as well as the focus position of both lenses.

4 Summary

We have in this paper presented a simple method for controlling the vergence mechanism of an agile camera head. The vergence and version processes have been totally separated, and may as such be controlled individually. The ideas suggested here have been motivated by primate vision.

The paper has presented initial experiments on controlling the camera head using accommodation and disparity cues. Currently the accommodation cue requires some improvement.

References

[1] Lynn Abbott and Narendra Ahuja. Surface reconstruction by dynamic integration of focus, camera vergence, and stereo. In *Proceedings of the Second International Conference on Computer Vision*, pages 532–543, 1988.

[2] Claus Siggaard Andersen, Jan Juul Sørensen, and Henrik Iskov Christensen. An analysis of three depth recovery techniques. In *Proceedings of the Seventh Scandinavian Conference on Image Analysis*, pages 66–77, 1991.

[3] Heinrich Bülthoff and Hanspeter A. Mallot. Interaction of different modules in perception. In *Proceedings of the First International Conference on Computer Vision*, pages 295–305, 1987.

[4] R.H.S Carpenter. *Movements of the Eyes.* Pion, London, 1988.

[5] David J. Coombs and Christopher M. Brown. Cooperative gaze holding in binocular vision. *IEEE Control Systems*, pages 24–33, June 1991.

[6] Subhodev Das and Narendra Ahuja. Integrating multiresolution image acquisition and coarse-to-fine surface reconstruction from stereo. In *Proceedings of Workshop on Interpretation of 3D Scenes*, pages 9–15, 1989.

[7] Erik Paul Krotkov. *Active Computer Vision by Cooperative Focus and Stereo.* Springer Verlag, New York, 1989.

[8] Thomas J. Olson and David J. Coombs. Real-time vergence control for binocular robots. *International Journal of Computer Vision*, 7(1), 1991.

[9] Kourosh Pahlavan, Thomas Uhlin, and Jan-Olof Eklundh. Dynamic fixation. In *Proceedings of the Fourth International Conference on Computer Vision*, pages 412–419. IEEE, 1993.

[10] Yehezkel Yeshurun and Eric L. Schwartz. Cepstral filtering on a columnar image architechture: A fast algorithm for binocular stereo segmentation. *IEEE Transactions on Pattern Analysis and Machine Inteligence*, 11(7):759–767, 1989.