MVA'94 IAPR Workshop on Machine Vision Applications Dec. 13-15, 1994, Kawasaki

Thorough ZDF-Based Localization for Binocular Tracking

N. Kita †S. Rougeaux Y. Kuniyoshi S. Sakane

Autonomous Systems Section, Electrotechnical Laboratory
1-1-4 Umezono,Tsukuba,Ibaraki 305,JAPAN

†Institut d'Informatique d'Entreprise
18 Allee Jean Rostand,91000 Evry,FRANCE

Abstract

We present an efficient method to fixate a binocular gaze point on a target object moving around in a complicated environment. Assuming the target is in the vicinity of the gaze point, the image features of the target can be isolated by using zero disparity feature from others which suffer some disparities. This assumption is proper while successful binocular tracking, and the zero disparity filtering (ZDF) is computationally very cheap enough to realize quick visual feedback on an ordinal image processing hardware. The position estimation obtained from such isolated features, however, is restricted in zero disparity fields, resulting in no depth cue. In order to get depth information simultaneously by ZDF processing, we introduce the novel localization technique based on the idea of "virtual horopter". The proposed method is implemented on our active vision system and the total performance of binocular tracking is demonstrated by actual tracking experiments.

Keywords : *Active vision, Binocular tracking, Zero disparity, Horopter, Gaze control, Fixation*

1 Introduction

In the field of computer vision research, various problems have been approached by assuming the visual data obtained from fixed cameras. Recently, it was claimed that various early vision processing, such as shape from shading, structure from motion, optical flow detection and so on, become more simple and robust by utilizing controlled motion of cameras [1]. Furthermore from more general points of view, some researchers summarized and partially demonstrated the benefits introduced by the dynamic control of camera parameters including both geometric and optic ones [2, 3, 4]. Following these precursors, lots of research groups in both field of computer vision and robotics have been trying to construct agile camera systems which should be visually controlled. They are also trying to implement reflex level visual behavior on their active vision system, which are promisingly composing higher level visual behavior for purposive application such that object recognition, manipulation, navigation and so on. Such underlying visual behaviors are desired to be realized with as possible as less computational power and time.

Among such fundamental visual behaviors, binocular tracking, which control the gaze position onto a moving or stationary target from the moving or stationary platform as the image of the target falls onto the fovea of the image plane, is most important. The gaze control for binocular tracking is typically carried out as follows; localizing the target on both images taken by right and left cameras, estimating the 3D position of the target and orienting right and left optical axes to the estimated position independently. Visual localization in each image is the main subject for completing this loop, because the 3D position estimation after localization is easily done by simple triangulation method and motor commands for camera orientation can be derived by simple inverse kinematic. Usually localization based on an object model is computationally expensive except such particular case that an object can be discriminated by simple visual features like color. Then most of practical tracking systems are using precategorical features such as stereoscopic disparity for localizing a target which need not recognize the target [5]. Especially zero disparity feature plays important roll for the selectivity in the direction of depth.

In the next section, we shows the property of stereoscopic disparity and clarify that the combination with image window forms a 3D window. In the section 3, the position estimation method only based on zero dis-

parity features is proposed, and in the section 4, the control strategy to track a target object in the complicated environment is explained. The implementation of the method on our active vision system and experiments on tracking an object moving in real world are shown. After briefly mentioning about the applications of our binocular tracking system, we finally summarize.

2 Isolating by ZDF and image window

We assume a camera system as shown in fig.1. It has two cameras, whose optic axes are in the same plane, and has two degree-of-freedom (DOF) for driving their pan angles independently. The position of the gaze point, which is in fact the intersection of both optical axes, is controlled within the plane involving their optical axes by both pan angles.

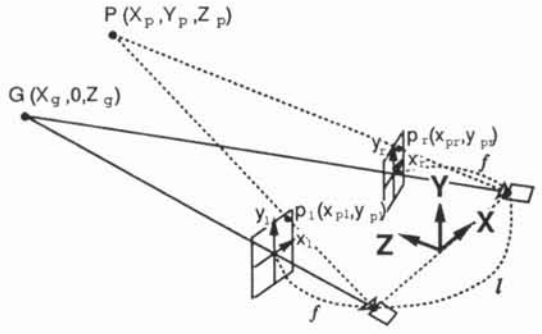

Figure 1: The model of a converging stereo system.

Under the definition of the world coordinate system, XYZ, and left and right image coordinate systems, x_l, y_l and x_r, y_r, as shown in fig.1, when the gaze point exists at $(X_g, 0, Z_g)$, the point in 3D space (X_p, Y_p, Z_p) is projected onto the image points, $p_l(x_{pl}, y_{pl})$ and $p_r(x_{pr}, y_{pr})$, on each image plane. Here,

$$x_{pl} = f\frac{(X_p + l/2)\cos\theta_l + Z_p \sin\theta_l}{-(X_p + l/2)\sin\theta_l + Z_p\cos\theta_l},$$

$$y_{pl} = f\frac{Y_p}{-(X_p + l/2)\sin\theta_l + Z_p\cos\theta_l},$$

$$x_{pr} = f\frac{(X_p - l/2)\cos\theta_r + Z_p \sin\theta_r}{-(X_p - l/2)\sin\theta_r + Z_p\cos\theta_r},$$

$$y_{pr} = f\frac{Y_p}{-(X_p - l/2)\sin\theta_r + Z_p\cos\theta_r}.$$

θ_l、 θ_r are the angles between Z axis and each of left and right optical axes, and

$$\cos\theta_l = \frac{Z_g}{\sqrt{(l/2 + X_g)^2 + Z_g^2}},\ \sin\theta_l = \frac{-(l/2 + X_g)}{\sqrt{(l/2 + X_g)^2 + Z_g^2}},$$

$$\cos\theta_r = \frac{Z_g}{\sqrt{(l/2 - X_g)^2 + Z_g^2}},\ \sin\theta_r = \frac{(l/2 - X_g)}{\sqrt{(l/2 - X_g)^2 + Z_g^2}}$$

The 3D coordinates that have zero horizontal disparity, which makes a cylindrical shape, is derived by solving the equation, $x_{pl} - x_{pr} = 0$, that is

$$X_p^2 + (Z_p - \frac{l}{2}\tan^{-1}(\theta_r - \theta_l))^2 = (\frac{l}{2})^2 + (\frac{l}{2}\tan^{-1}(\theta_r - \theta_l))^2.$$

The horizontal section by XZ plane is a circle which includes the gaze point and both optical centers. This is so called horopter. By the similar procedure, the locus of the points which have some amount of horizontal disparity can be derived. Figure 2 shows such loci of the points that have horizontal disparities of 0 and $\pm 10n (n = 1 \sim 5)$ pixels in the case that the focal length f is 6 (unit is mm), baseline length l is 200, the gaze position is $(200, 0, 400)$, and the image plane, whose size is 6.48×4.82, is digitized to 512×480 pixels.

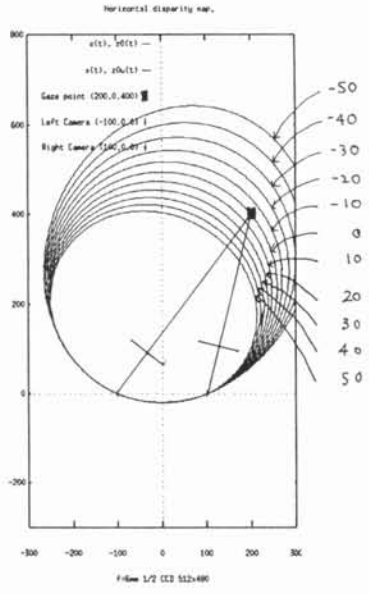

Figure 2: Horizontal disparity map.

As shown in the figure, at the vicinity of the gaze point the horizontal disparity is increasing in the direction nearly perpendicular to both of image planes, that is the depth direction. The object inside of the region surrounded by the interrupted lines in the fig.3

can be discriminated by the condition that the horizontal disparity is close to zero. Consequently, the combinatorial use with image window, which suppress the object outside of the region surrounded by the dotted lines in the fig.3, forms a 3D window which picks up only the objects in the space surrounded by the solid lines in the figure 3.

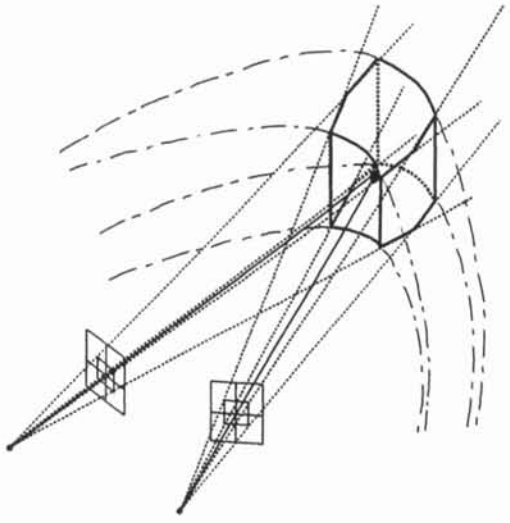

Figure 3: The possible space selected by the joint use of the horizontal disparity and the image window.

This isolating technique, which is called as zero disparity filtering (ZDF), is easily realized by the simple image processing as follows. First, vertical edges within the image windows are extracted from both left and right images, next blurred and then binalized by a predefined threshold, and finally the logical AND operation are performed between right and left binary images (fig.4).

To exploit ZDF to extract the target object, which resides near the gaze point during binocular tracking, we set the size of the 3D window so as to involve the area, where the target is expected to exist, according to the speed, size and the distance from the cameras. Consequently, we can pick up the image features of the target as the output of ZDF. But the output does not give the position information in the depth direction. Moreover, the image features passing through ZDF are not only ones of the target. In the next section, we discuss the way to estimate the target position based on the output of ZDF.

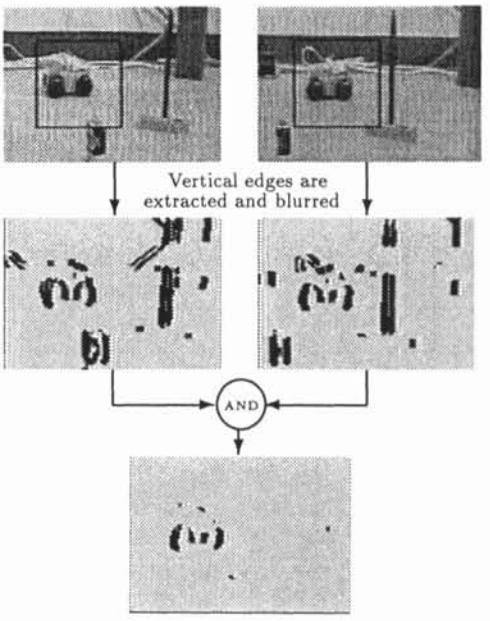

Figure 4: ZDF processing.

3 Position estimation by virtual horopter

Though we can not derive the depth position of the target directly from the outputs of ZDF, we can estimate on which locus in the fig.2 the target resides from the width of the edges in ZDF outputs. For example, if the width of a edge is equal to the blurred width "w", the edge is just on the horopter, and if the width of a edge is "$w - \Delta w$", the edge resides on either of two loci corresponding to the horizontal disparity of "$\pm\Delta w$" respectively.

In order to decide on which loci of $\pm\Delta w$ the target exists, we examine ZDF outputs after moving the horopter to the loci. If actually rotating cameras to move the horopter, the examination process need much time. To cope with it, we introduce the novel idea, "virtual horopter". The virtual horopter is the horopter generated by shifting horizontally the right image. As shown in figure 5, small shifts (s pixel) of the right image are almost equivalent to small *virtual* rotations ($\Delta\theta_r$) of the right camera. Here,

$$\Delta\theta_r = \tan^{-1}(s/f_x),$$

$$f_x = f \times h_{pixel}/h_{width},$$

f is the focal length, h_{pixel} is the pixel width of the images, and h_{width} is the horizontal length of the image planes.

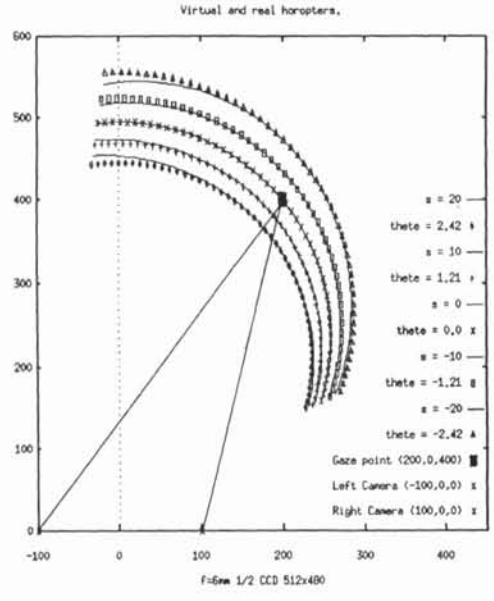

Figure 5: *Virtual horopters (solid lines) and the corresponding real horopters (dotted lines).*

Thus we can substitute the actual rotation of cameras by image shifts. Consequently, after ZDF processing at the three different horopters, the target is estimated as existing on the horopter that produces the most outputs.

4 Control strategy

If there is only one vertical edge belonging to a target object in ZDF output, the position of the target is correctly estimated by the proposed method. But a target object usually shows more than one vertical edges in ZDF outputs. Then we estimate the target position using the center of gravity of ZDF outputs. Regarding the depth estimation, instead of using the width of a edge, we use the average width of all edges in a ZDF output. Concretely, we count the total number "N" of pixel in the ZDF output and convert this into a pixel shift "s" by the following equation.

$$s = w \times (1 - \frac{N}{N_{max}})$$

Here, N_{max} is the pixel number in the ideal case, that is when the target is on the horopter, and N_{max} is actually the pixel number on the ZDF output at the beginning of the tracking. As demonstrated in the experiments shown in the section 5, this estimation

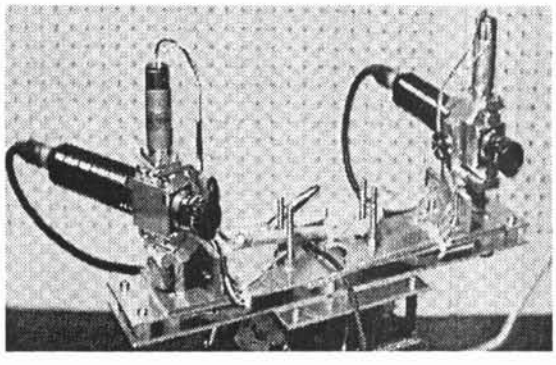

Figure 6: *The gaze platform.*

strategy works well in the environment, whose background is not uniform and where several other objects move around.

In a ZDF output, however, not only edges originating a target object, but also edges originating the other objects or two different objects, which make their projections coincidentally at the same coordinates in left and right images respectively, appears. Moreover a target object may be occluded. In such cases, the position estimation suffers from errors, and if the error exceed an amount, the tracking fails. Empirically this happens in the following situations. (1) There are dense vertical edges in the background. (2) Another object which shows fairly large number of vertical edges comes into the 3D window area. (3) A target object is largely occluded. The tracking failure in the cases of (1) and (2) can be avoided by using only edges originating the target object. We can extract and accumulate the knowledge about the target object while tracking, and it can be used to identify the target. Though such identification procedure needs fairly complicated computation, it is necessary only when the situation, (1) or (2), occurs, and the occurrence of the situations can be judged by simply examining the horizontal distribution of ZDF output, which is derived as a byproduct during the calculation of the center of gravity. Regarding the situation (3), position estimation of an occluded target is impossible, then to prevent the failure of the tracking, we halt the tracking motion while the target is occluded. The occurrence of the occlusion is judged by examining the total pixel number of ZDF outputs.

5 Active vision system

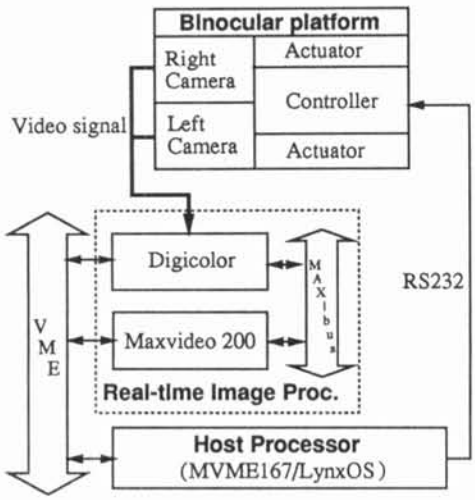

Figure 7: *Overview of our active vision system.*

We have built a prototype of gaze platform (fig.6), which has two degrees of freedom corresponding to the two rotations of the cameras around the vertical axis. It is equipped with DC motors and potentiometers for pan angle range measurements. The stereo image, which are acquired by two CCD cameras, are digitized at the frame grabber (DigiColor) and sent to a pipeline image processor (Maxvideo200) (fig.7). ZDF is performed at Maxvideo200 and the position estimation at host processor (MVME167). The host processor also manages the visual feedback loop based on the position estimation of a target. The feedback loop runs at 30 Hz with about 90 $msec$ delay.

6 Experiments

In the following experiments, the base length was set to 200mm and 6mm focal length lenses were used. In the first experiment, a turn table was set in front of the gaze platform at the distance of 1 meter and a target was put on the turn table at 30 cm from the center. Figure 8 shows motor response while tracking the target when the turn table was rotating at 0.4 Hz. Error means the angle difference between the optimal (target) position and the real positions of the cameras. This shows that the system tracks a target with the latency of 90 msec. After many trials with various rotating speeds of the turn table and distances between the target and the center of the table, the speed limits which the system can track were about $50^o/sec$ for the movements along the horopter and about $15^o/sec$ for the movements across the horopter, in terms of the angular speed of only right camera rotation. The reason of the former limitation is that the maximum speed of camera rotations are $50^o/sec$. The latter limitation corresponds to the 4 pixel shift in one frame time, which is about 70 % of the blurred edge width.

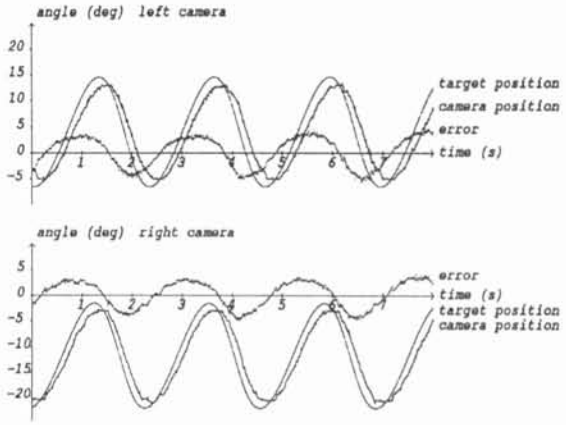

Figure 8: *Motor response while tracking a target on a turn table.*

Various shaped objects, not only objects consisting of vertical edges like a pencil put vertically, but objects like a miniature car or a doughnut, could be tracked with several distractors (figure 9). This demonstrates that the ZDF method based on vertical edges has fairly well generality about the target shape and capable to discriminate the target from backgrounds or distractors. The system sometimes halted its tracking behavior in the following cases; when a lot of vertical edges besides ones of the target appeared in both fields of views of right and left cameras, the other moving object came close to the target object or the target object was occluded by the other objects. The tracking, however, was restarted when the zero disparity features of the target became dominant over the ZDF output again.

7 Application of binocular tracking system

Our binocular tracking system has several functions besides tracking. It can measure the relative position of a target while tracking it, fixate the gaze point to a stationary object, and check the existence of an object in the particular 3D space by using its 3D window function. We have already applied these functions to

control cooperative behaviors of multi mobile robots [6]. Typical one of such cooperative behaviors is the posing behavior, in which a mobile robot mounting the binocular tracking system autonomously follows another mobile robot while keeping a constant distance.

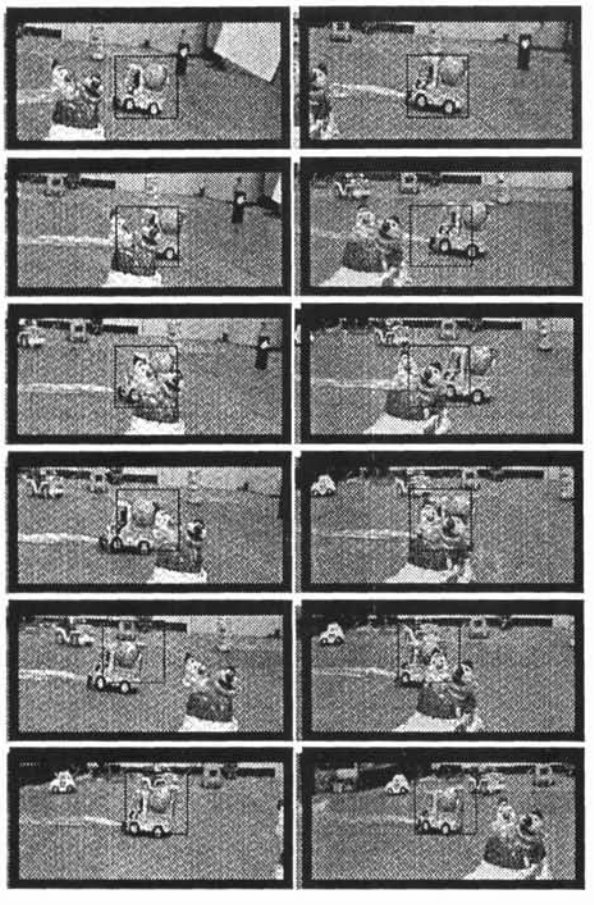

Figure 9: *The left and right camera views while tracking a toy car.*

8 Conclusion

We proposed a simple method which localizes a target object based on zero disparity features in order to hold the binocular gaze point on the target. To localize the depth position of a target from zero disparity features, which has no explicit information about depth, we newly introduced the concept of *virtual horopter*. This enabled the 3D position estimation only with zero disparity filtering, resulting in much simpler and computationally cheaper method than the conventional method [7].

We implemented the proposed method on the prototype of active vision system, and carried out experiments of binocular tracking for various kinds of objects in different environments. These experiments demonstrated that the proposed method well worked for fairly complicated situation. We also showed the direction of the further improvements of binocular tracking method.

In the current situation, the system requires that the target object is at the gaze point for each initialization of the tracking task. We already started to integrate with a target acquisition algorithm based on optical flow. This will make the binocular tracking system more applicable for various kinds of visual tasks.

References

[1] J. Aloimonos, I. Weiss and A. Bandyopadhyay: "Active Vision" , *Proc. of First ICCV*, 552–573, 1987.

[2] R. Bajcsy: "Active Perception vs. Passive Perception" , *Proc. of IEEE Workshop on Computer Vision*, 55–62, 1985.

[3] Dana H. Ballard: "Animate Vision" , *Artificial Intelligence*, no.48, 57–86, 1991.

[4] J. K. Tsotsos: "Active vs. Passive Visual Search: Which is more Efficient?" , *Tech. Rep. on Univ. of Toronto*, RBCV-TR-90-34, 1990.

[5] A. Maki, T. Uhlin and J.-O Eklundh: "Phase-based disparity estimation in binocular tracking" , –, 1993.

[6] Y. Kuniyoshi, S. Rougeaux, M. Ishii, N. Kita, S. Sakane and M. Kakikura: "Cooperation by observation –The framework and basic task patterns–" , *Int. Conf. on Robotics and Automation*, 1994.

[7] D. J. Coombs and C. M. Brown: "Real-time binocular smooth pursuit" , *Int. Journal of Computer Vision*, vol.11, no.2, 1993.