Real Time Pursuit and Vergence Control with an Active Binocular Head

K. Daniilidis, M. Hansen, G. Sommer

Institut für Informatik und Prakt. Mathematik
Christian-Albrechts Universität Kiel
Preusserstr. 1-9, 24105 Kiel, email:{kd,mha,gs}@informatik.uni-kiel.de

Abstract: This article is concerned with the design and implementation of a system for real time tracking of a moving object and binocular vergence control for depth estimation. Object detection relies only on the image motion without making any a priori assumptions about the object form. Using only the first spatiotemporal image derivatives subtraction of the normal optical flow induced by camera motion yields the object image motion. On the other hand, both the left and the right image are filtered hierarchically with Gabor functions. The phase difference of the responses yields a disparity map. The disparity in the center is the reference signal for vergence control of the binocular head. Both behaviors are implemented in parallel and the cycle rate achieved is 25 Hz.

1 Introduction

Traditional computer vision methodology regarded the visual system as a passive observer whose goal is the recovery of a complete description of the world. This approach led to systems unable to interact in a fast and stable way with a dynamically changing environment. The new paradigm of active behavioral vision showed that the ability to control the mechanical degrees of freedom during image acquisition as well as the behavior dependent selectivity in data processing facilitate more stable and real-time reactions in navigation and manipulation tasks.

The most evident reason for object pursuing is the limited field of view available by CCD cameras. The two degrees of freedom of panning and tilting enable keeping a moving object of interest in view for a longer time interval. Even if we had a sensor with 180 degrees field of view it would not be computationally possible to process every part of the field of view in the same detail. On the other hand, vergence control keeps the disparity magnitude bounded avoiding, thus, a computationally expensive search in large regions. Working with small disparities allows the use of filters with smaller support and reduces the computation time. Potential applications for the presented system are in the field of surveillance in indoor or outdoor scenes. The advantages are not only in the motion detection but mainly in the capability of keeping an intruder inside the field of view. Another application is in automatic video recording and video teleconferencing. The camera automatically tracks the acting or speaking person so that it always remains in the center of the field of view. In manufacturing or recycling environments, an active camera can track and estimate the depth to objects on the conveyor-belt so that they are recognized and grasped without stopping the belt. New directions are opened if such an active camera platform is mounted on an autonomous vehicle. As we will show in the results, vergence control supports scene exploration and the building of an environmental map.

The novelty of this approach is in the achievement of a video rate tracking and vergence control using sound image processing techniques. The performance of 25 Hz with a latency of about 100ms classifies our system together with the systems of Oxford and Stockholm among the fastest systems worldwide. Novel is also the design of the derivative and Gabor filters with respect to the limited support and accuracy in the given pipeline architecture. We demonstrate that real time implementation is not achieved by introducing heuristics but by systematic filter design.

As pursuing is one of the basic capabilities of an active vision system most of the research groups possessing a camera platform have reported results. The Oxford surveillance system [6] uses data from the motor encoders to compute and subtract the camera motion induced flow. It runs in 25 Hz with processing latency of about 110 ms. Camera behavior is modeled as either saccadic or pursuit motion and a finite state automaton controls the switching between the two reactions. The KTH-Stockholm system [10] computes the ego motion of the camera by fitting an affine flow model in the entire image. It is the only approach claiming pursuit in presence of arbitrary observer motion and not only pure rotation as assumed by the rest of the algorithms. However, this global affinity assumption is valid only if the object occupies a minor fraction of the field of view which is not a realistic assumption. Elimination of the flow due to known camera rotation is also applied by Murray and Basu [5]. The background motion is compensated by shifting the images. Then large image differences are combined with high image gradients to give a binary image. No real time implementation results are reported.

Regarding vergence control different approaches are developed. Olson and Coombs [7] used a cepstral filter for disparity estimation, and developed a real time vergence control with a servo rate of 10 Hz. Closer to our approach in disparity estimation was the work of Theimer & Mallot [9]. They also used a phase-based approach with Gabor filters on sub-sampled images with a rate of 1 Hz on common hardware. Westelius et al. [11] developed a vergence control based on phase differences. To get stable results they additionally computed the disparity from a pair of edge images. Uhlin et al. [10] also implemented a vergence control with phase-based disparity estimation and achieved the servo rate of 25 Hz.

2 Monocular pursuit of a moving object

A moving object in the image is defined as the locus of points with high image gradient whose image motion is substantially different from the camera induced image motion. We exploit the fact that the camera induced optical flow $\boldsymbol{u_c}$ is pure rotational

$$\boldsymbol{u_c} = \begin{pmatrix} x_c y_c & -(1+x_c^2) & y_c \\ (1+y_c^2) & -x_c y_c & -x_c \end{pmatrix} \boldsymbol{\omega}$$

where (x_c, y_c) are the camera coordinates and $\boldsymbol{\omega}$ is the angular velocity computed from the angle readings of the axis encoders as follows.

The binocular camera mount[1] used in our system has four mechanical degrees of freedom: the pan angle χ of the neck, the tilt angle ϕ, and two vergence angle θ_l und θ_r for left and right, respectively (Fig. 1). The stereo basis is denoted by B.

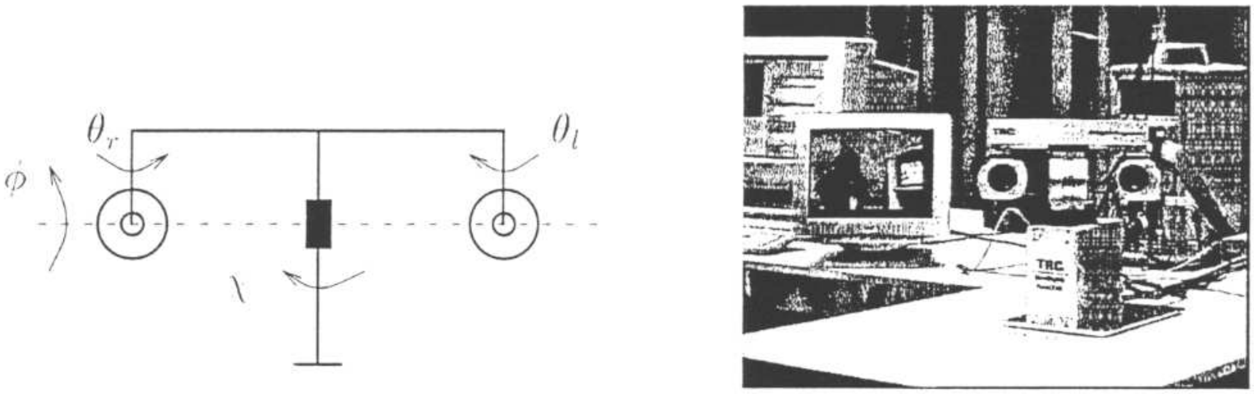

Abb. 1. The four degrees of freedom of the camera platform (left) and how it looks like (right).

Let $R(t) = R_{\phi(t)} R_{\theta(t)}$ be the time varying rotation of the camera coordinate system and Ω the skew-symmetric tensor of the angular velocity. Then we have $\dot{R}(t) = R(t)\Omega$ and the angular velocity with respect to the moving coordinate system reads $\boldsymbol{\omega} = \left(\dot{\phi}\cos\theta \;\; \dot{\theta} \;\; \dot{\phi}\sin\theta \right)^T$. We assume the Brightness Change Constraint Equation $g_x u + g_y v + g_t = 0$ with g_x, g_y and g_t the spatiotemporal derivatives of the grayvalue function. From this equation we can compute only the normal flow - the projection of optical flow in the direction of the image gradient (g_x, g_y). The difference between the normal flow u_{c_n} induced by camera motion and the observed normal flow u_n

$$u_{c_n} - u_n = \frac{g_x u_c + g_y v_c}{\sqrt{g_x^2 + g_y^2}} + \frac{g_t}{\sqrt{g_x^2 + g_y^2}}$$

is the normal flow induced by the object motion. It turns out that we can test the existence of object image motion without the computation of optical flow. The sufficient conditions are that the object motion has a component parallel to the image gradient and the image gradient is sufficiently large. We can thus avoid the computation of full optical flow which would require the solution of at least a linear system for every pixel. Three thresholds are applied: the first for the difference between observed and camera normal flow, the second for the magnitude of the image gradient, and the third for the area of the points

[1] Consisting of the TRC BiSight Vergence Head and the TRC UniSight Pan/Tilt Base

satisfying the first two conditions. The object position is given as the centroid of the detected area.

2.1 Real time spatiotemporal filtering

Special effort was given to the choice of filters suitable for the used pipeline-processor [2] so that the frequency domain specifications are satisfied without violating the real time requirements. Whereas up to 8 × 8 FIR-kernels can be convolved with the image with processing rate of 20 MHz the temporal filtering must be carried out by delaying the images in the visual memory. We chose IIR filtering for the computation of the temporal derivatives since its computation requires less memory than temporal FIR filtering for the same effective time lag.

The temporal lowpass filter chosen is the discrete version of the exponential [3]

$$E(t) = \begin{cases} \tau e^{-t\tau} & t \geq 0 \\ 0 & t < 0. \end{cases}$$

If $E_n(t)$ is the n-th order exponential filter ($n \geq 2$) its derivative reads

$$\frac{dE_n(t)}{dt} = \tau(E_{n-1}(t) - E_n(t)).$$

The discrete recursive implementation for the second order derivative filter reads

$$h_1(k) + rh_1(k-1) = q(g(k) + g(k-1))$$
$$h_2(k) + rh_2(k-1) = q(h_1(k) + h_1(k-1)) \qquad g_t(k) = \tau(h_1(k) - h_2(k)),$$

where $g(k)$ is the input image, $h_1(k)$ and $h_2(k)$ are the lowpass responses of first and second order, respectively, and $g_t(k)$ is the derivative response. We note, that the lowpass response is used to smooth temporally the spatial derivatives. The spatial FIR-kernels are binomial approximations to the first derivatives of the Gaussian function.

2.2 Estimation and Control

The control goal of pursuing is to hold the gaze as close as possible to the projection of a moving object. Output measurements are the absolute position of the object denoted by $\boldsymbol{o}$ obtained from the centroid in the image and the angle readings. Let $\boldsymbol{v}$ and $\boldsymbol{a}$ be the velocity and acceleration of the object, $\boldsymbol{c}$ be the absolute position of the optical axis, and $\Delta\boldsymbol{u}(k)$ the incremental correction in the camera position. The state is described by the vector $\boldsymbol{s} = \begin{pmatrix} \boldsymbol{c}^T & \boldsymbol{o}^T & \boldsymbol{v}^T & \boldsymbol{a}^T \end{pmatrix}^T$. A motion model of constant acceleration and a linear control function $\Delta\boldsymbol{u}(k) = -K\hat{\boldsymbol{s}}(k)$ with $\hat{\boldsymbol{s}}$ an estimate of the state enables the use of the separation principle stating that optimal control can be obtained by combining the optimum deterministic control with the optimal stochastic observer. The minimization of the difference $||\boldsymbol{o} - \boldsymbol{c}||$ between object and camera position in the reference coordinate system can be modeled as a Linear Quadratic Regulator problem with the minimizing cost function $\sum_{k=0}^{N} \boldsymbol{s}^T(k)Q\boldsymbol{s}(k)$ where Q is a symmetric matrix with $Q_{11} = 1$,

[2] Datacube MaxVideo 200 board

$Q_{12} = Q_{21} = -1, Q_{22} = 1$ and the rest of its elements zero. In steady state modus a constant control gain K is assumed resulting in an algebraic Ricatti equation with the simple solution that input camera position should be equal the predicted position of the object. One of the crucial problems in vision based closed loop control is how to tackle the delays introduced by a processing time longer than a cycle time. We emphasize here that the delay in our system is an estimator delay. The normal flow detected after frame k concerns the instantaneous velocity at frame $k-1$ due to the mode of the IIR temporal filter. At time $k-1$ the encoder is also asked to give the angle values of the motors. To the delay amount of one frame we must add the processing time so that we have the complete latency between motion event and onset of steered motion.

Concerning optimal estimation we also assume steady state modus obtaining a stationary Kalman Filter with constant gains. The special case of a second order plant yields the well known α-β-γ-Filter with update equation

$$\hat{\boldsymbol{s}}^{+}(k+1) = \hat{\boldsymbol{s}}^{+}(k) + (\,\alpha \;\, \beta/\Delta t \;\, \gamma/\Delta t^2\,)^T(\boldsymbol{m}(k+1) - \boldsymbol{m}^{-}(k+1)),$$

where $\boldsymbol{s}^{+}$ is the state after updating and $\boldsymbol{m}^{-}$ is the predicted measurement. The gain coefficients α, β and γ are functions of the target maneuvering index λ. This maneuvering index is equal to the ratio of plant noise covariance and measurement noise covariance. The lower is the maneuvering index the higher is our confidence in the motion model resulting to a smoother trajectory. The higher is the maneuvering index the higher is the reliability of our measurement resulting to a close tracking of the measurements which may be very jaggy. This behaviour was thoroughly studied in [1].

We proceed with a real experiment. In Fig. 2 the system is tracking a Tetrapak moving from right to left. The images in Fig. 2 are chosen out of 20 frames saved "on the fly" during a time of 8s. The centroid of the detected motion area is marked with a cross. We show the tracking error by drawing the trajectory of the centroid in the image as well as the control values for the tilt and the vergence angle, ϕ and θ for the entire time interval of 8s. Although the target might move smoothly the orbit of the centroid depends on the distribution of the detected points in the motion area. Therefore, it is corrupted with an error of very high measurement variance. Allowing a high maneuvering index which enables close tracking would result in a extreme jaggy motion of the camera. The estimator would forget the motion model and yield an orbit as irregular as the centroid motion. Therefore, we decrease the maneuvering index to 0.01 and obtain as expected a much higher pixel error. Only a post processing of the binary images could improve the position of the detected centroid. The small size of the target enables a relatively small pixel error (the target is always observed left from the center). Because the centroid variation is only in the vertical direction - due to the rod holding the target - the tilt angle changes irregularly. The average angular velocity is 8.5 deg/s. The reader is referred to [1] for numerous real and synthetic experiments.

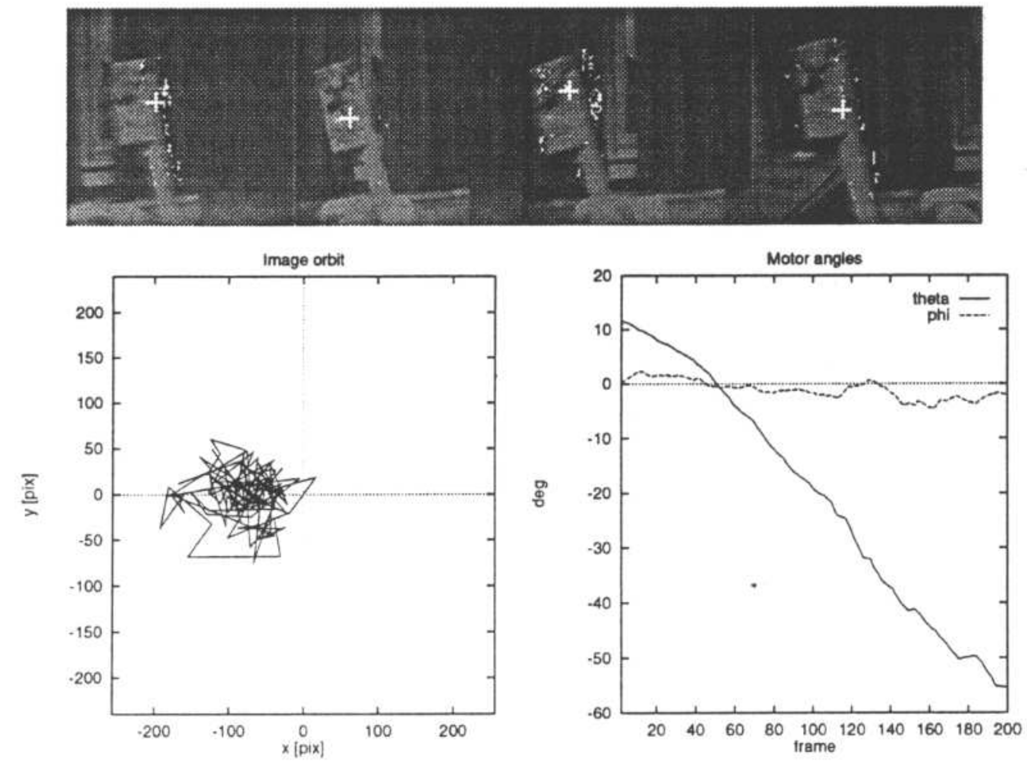

Abb. 2. Four frames recorded while the camera is pursuing a Tetrapak moving from right to left. The pixel error (bottom left) shows that the camera remains behind the target and the vergence change (bottom right) shows the turning of the camera from right to left with an average angular velocity of 8.5 deg/s.

3 Phase-based Disparity Estimation for Vergence Control

The idea of a phase-based approach is to implicitly solve the correspondence problem. Without explicit feature extraction these approaches can be described similarly by a local correlation of bandpass-filtered images. The local phase response contains the information of the spatial position of the matched structure. According to the Fourier shift-theorem

$$f(x) \;\circ\!\!-\!\!\!-\!\!\!-\!\!\bullet\; F(\omega) \quad f(x+D) \;\circ\!\!-\!\!\!-\!\!\!-\!\!\bullet\; F(\omega)e^{i\omega D},$$

a global spatial shift D of a signal $f(x)$ can be detected as a phase shift in the Fourier spectrum. Extracting the local phase in both images of a stereo pair with complex filters like Gabor filters leads to a direct computation of local disparity. Fleet [2] and Sanger [8] have employed phase-based approaches to recover disparity information with complex Gabor filters on different scales. The spatial shift $D(x) = \frac{\Delta\Phi(x)}{\omega}$ is computed in the *constant frequency model* from local phase difference [2]:

$$\Delta\Phi(x) = \phi_l(x) - \phi_r(x). \tag{1}$$

The phases are denoted $\phi_r(x)$ for the right and $\phi_l(x)$ for the left image.

Our Approach for a fast real-time algorithm [4] is influenced by the given hardware to obtain real time performance. Small filters and a simple algorithm can perform a high clock rate. We developed such a simple algorithm based on the theoretical principles of the phase-based approach.

3.1 Filter design

Our Gabor filters with an odd size of 7x7 regard the following four constraints, noticed in [11]:

a. *No DC component* to get an optimal phase behaviour.

b. *Suppression of wrap around* of the phase for maximizing the measurable disparity.

c. *Monotonous phase* to assure the one to one relation between phase difference and disparity.

d. *Small support* to get low computational costs.

To get no wrap around and to have a maximum measurable horizontal disparity related to the filter size a wavelength $\lambda = 2\pi/\omega_x = 6$ pixel is optimal.

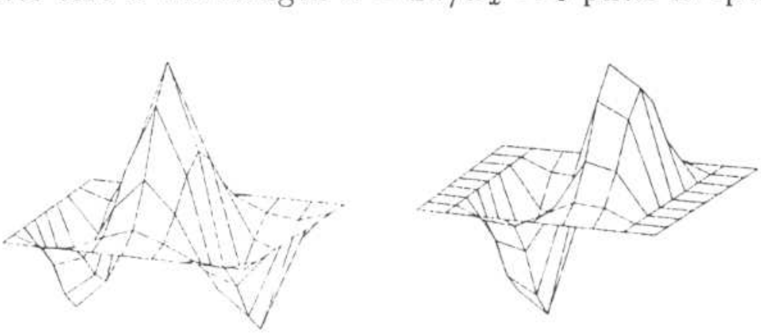

Abb. 3. Odd and Even Gabor filter 7x7.

Without a consistency check of the measured phases in the left and the right image, disparity estimation can produce arbitrary results. Our method to check the stability of phase information based on thresholding the magnitude of the filter responses. First, in each image the magnitudes are thresholded (by 20 % of the maximum magnitude) and second, the sum of magnitudes of both images are calculated and thresholded again (by 40 % of the sum maximum magnitude). The estimated phase difference at image pixel x is called stable if these two constraints are fulfilled. This map of stable phase differences is our *confidence map* $c(x)$.

The small filter size demands a strategy to deal with larger disparities in stereo images. Our approach is to compute a Gaussian pyramid $f_i(x)$ by approximating the Gaussian filter by a 7x7 binomial filter B. Sub-sampling S reduced the image resolution from 512x512 at the finest level to 16x16 at the coarsest level. This results in a maximal measurable disparity of ±192 pixel at the coarsest level.

$$f_g(x)_i = G(x, \sigma, \omega) * f_i(x) \quad f_i(x) = S(B * f_{i-1}(x)).$$

4 Vergence Control and one Application

The camera mount (Fig. 1) has four mechanical degrees of freedom: pan angle, tilt angle ϕ, and two vergence angle θ_r and θ_l. The right camera has been declared as the dominating eye of our system. Pan, tilt and right vergence angle are controlled by a gaze-controller. Regarding vergence movement we only have to control the left vergence angle θ_l to reduce the horizontal stereo disparity D_c in the center of view. The disparity D_c is picked out from the center of the computed disparity map. The vergence control is designed as a feedback loop.

The estimated disparities D_c are compared with the reference signal $D_0 = 0$ in the case of convergence. The left vergence angle θ_l is controlled by the PD-controller because of its robust behaviour in real time applications. The offsets $\Delta\theta_l$ is defined by the PD-control law:

$$\Delta\theta_l = K_p D_c + K_d \dot{D}_c \tag{2}$$

The controller gains K_p and K_d are tuned by the Ziegler-Nichols method to have robust control and minimal settling time.

One application of combining vergence and gaze control is active depth estimation in an unknown area. The gaze controller has to move the gaze direction of the dominating right camera to interesting points in the world. These are edges and corners in the case of well structured areas. Then the left camera can fixate the same points by vergence control. After verging depth can be computed. We use the responses of our Gabor filters for controlling the right camera by choosing the local maxima of one confidence map (128x128) as gaze points. Normally a lab scene contains 10 - 25 selected local maxima in a view. After estimating the range of all points a new confidence map is computed. Then local maxima can be detected again at this new view until the chosen segment of the unknown area is explored.

4.1 Depth computation

In the case of convergence on a gaze point P we need the knowledge of the left and right vergence angles θ_l, θ_r to compute the depth of this point. Additionally the baseline of the stereo rig is known. We compute the depth in a cyclopean frame (see Fig. 4). The origin O is at half base line B. The cameras are verging on selected gaze point P. The gaze point is denoted by the angles (γ, ϕ),where ϕ is the tilt angle and $\gamma = \arctan(\sin(\theta_l - \theta_r)/(2\cos\theta_l \cos\theta_r))$.

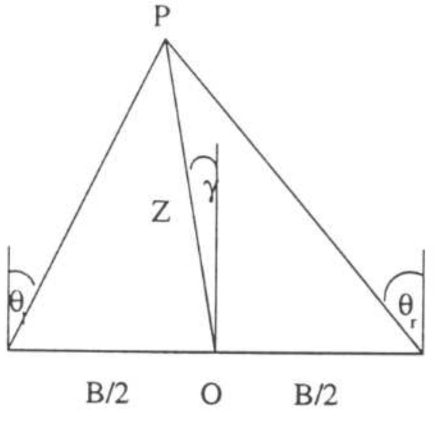

Abb. 4. Stereo geometry of our vision system at convergence at gaze point P.

We define the depth Z of the point P as the length of the line P to O. With trigonometric transforms the equation for $Z(\theta_l, \theta_r, B)$ follows:

$$Z^2 = B^2\left(\frac{\sin^2(\theta_l - \theta_r)}{4\sin^2(\theta_l + \theta_r)} + \frac{\cos^2\theta_r \cos^2\theta_l}{\sin^2(\theta_l + \theta_r)}\right). \tag{3}$$

We represent the depth map $Z(\gamma, \phi)$ with a resolution of $2\pi/360$ for γ and ϕ. The vergence angles have an intrinsic resolution of 0.006^o. This results in a theoretical error in depth estimation $\frac{\Delta Z}{Z}$ from 0.09% at $Z = 1.0$ m up to 0.9% at $Z = 10.0$m in the case of symmetric vergence.

5 Experiments and results

The example shows a typical view of our lab. We explore this scene with our system to get depth information. Fig. 5 shows a fly of four images (u.) and their resulting confidence maps at a resolution 128x128 (l.). The confidence maps are used for control the gaze direction (γ, ϕ). In this example the tilt angle $\phi = 0$ is hold constant. The range of the gaze angle γ is $-55^o ... 50^o$. The gaze controller selected 14 points from the four confidence maps. The white scan line (Fig 5.) is the area, where local maxima are detected. It represents the constraint $\phi = 0$ The gaze of the right camera is directed to each gaze point $(\gamma_i, 0)$. After verging the depth $Z(\gamma_i, 0)$ is computed. The tabular shows the estimated depth values (rounded in 0.05 m) and computed gaze directions (rounded in 1 degree).

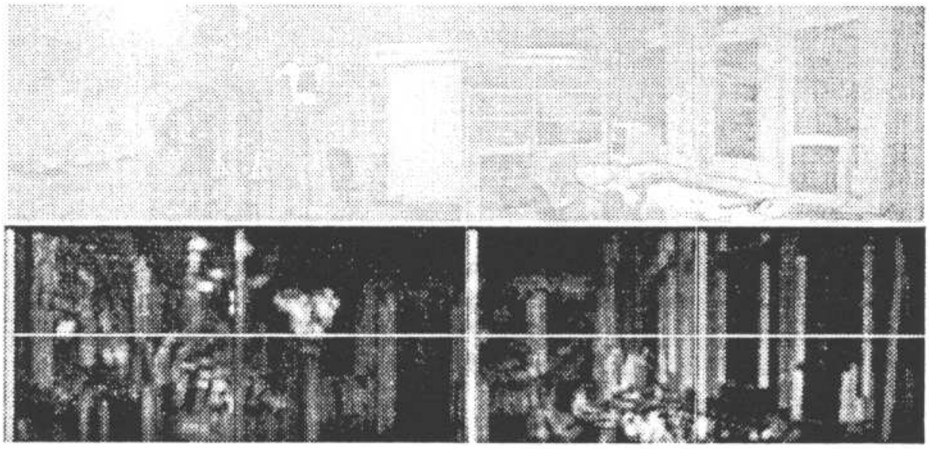

Abb. 5. The view of our lab which has to be explored. (l.) The confidence maps of the same images at resolution 128x128.

γ_i	-53	-50	-42	-38	-32	-27	-21	-12	-5	5	7	15	21	43
Z	3.30	3.65	3.95	4.30	4.70	5.30	6.15	5.95	5.90	6.00	6.10	4.05	4.50	2.30

Figure 6 shows a result of depth exploration. The depth $Z(\gamma, 0)$ is the radius of the polar figure. Linear interpolation is applied between gaze points $(\gamma_i, 0)$. In Fig. 6 it can be recognized the approximate rectangular outline of our lab. At the right side the windows, at the front side the open cupboards and the door and clipboard at the left wall have good structure, so that vergence control with Gabor filters was possible.

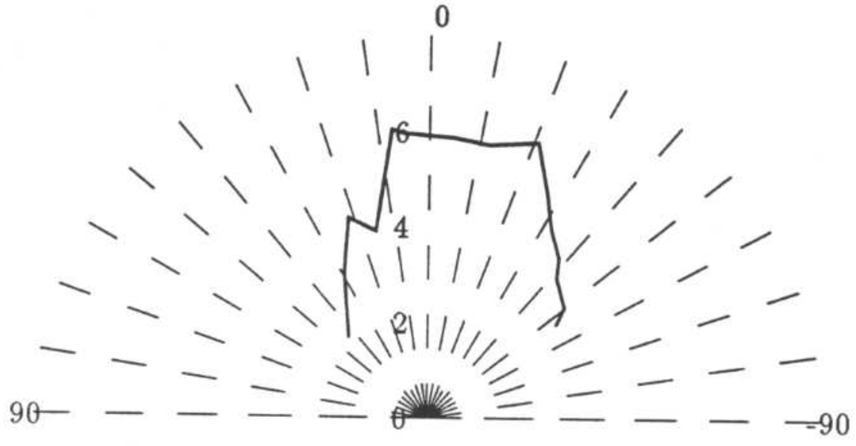

Abb. 6. Depth map $Z(\gamma_i, 0)$ of our lab. The gaze directions range from $-53... + 43°$.

References

1. K. Daniilidis, Ch. Krauss, M. Hansen, and G. Sommer. Real Time Tracking of Moving Objects with an Active Camera. Technical Report 9509, Inst. f. Inf. u. Prakt. Math., October 1995. submitted also to the *Real Time Imaging* Journal.
2. D. J. Fleet, A. D. Jepson, and M. R. M. Jenkin. Phase-Based Disparity Measurement. *CVGIP: Image Understanding*, 53(2), 3 1991.
3. D.J. Fleet and K. Langley. Recursive filters for optical flow. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17:61–67, 1995.
4. M.Hansen and G.Sommer. Real Time Vergence Control using Local Phase Differences. *Machine Graphics and Vision*, 5(1/2):51–63, 1996.
5. D. Murray and A. Basu. Motion tracking with an active camera. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16:449–459, 1994.
6. D.W. Murray, P.L. McLauchlan, I.D. Reid, and P.M. Sharkey. Reactions to peripheral image motion using a head/eye platform. In *Proc. Int. Conf. on Computer Vision*, pp. 403–411, Berlin, Germany, May 11-14, 1993.
7. T.J. Olsen and D.J. Coombs. Real-Time Vergence Control for Binocular Robots. *International Journal of Computer Vision*, 1:76–89, 1991.
8. T.D. Sanger. Stereo Disparity Computation Using Gabor Filters. *Biol. Cybernetics*, 59:405–418, 1988.
9. W.M. Theimer and H.A. Mallot. Phase-based binocular vergence control and depth reconstruction using active vision. *CVGIP: Image Understanding*, 60:343–358, 1994.
10. T. Uhlin, P. Nordlund, A. Maki, and J.A. Eklundh. Towards an Active Visual Observer. In *Proc. Int. Conf. on Computer Vision*, pp. 679–686. Boston, MA, June 20-23, 1995.
11. C.J. Westelius, H. Knutsson, J. Wiklund, and C.F. Westin. Phased-Based Disparity Estimation. In H.I.Christensen J.L. Crowley, editor, *Vision as Process*. Springer Verlag, Heidelberg, 1994.