Local stereoscopic depth estimation*

Kai-Oliver Ludwig, Heiko Neumann and Bernd Neumann

A method is discussed as to how a fixating binocular observer can recover local depth information with a single step computation avoiding the correspondence problem motivated from recent findings about the architecture of biological visual systems. Visual information is represented in the primate visual cortex (area 17, layer 4B) in a peculiar structure of alternating bands of left and right eye dominance. Recently, a number of computational algorithms based on this ocular stripe map architecture have been proposed. We investigated the cepstral filtering method of Yeshurun and Schwartz[1] for fast disparity computation because of its simplicity and robustness. Based on a systematic investigation and evaluation of the properties of the cepstral filter, some deficiencies are discussed, and improvements to the algorithm are presented. The introduction and brief review of the biological background may be skipped, if the reader is interested only in the technical aspects of the method. In summary, we consider Gaussian window functions for the extraction of local image patches superior to rectangular windows because specific configurations are avoided where additional maxima in the cepstral output may disturb the detection of the correct peak. We show experimentally that exact disparity estimates can still be obtained by the filter, even when one of the subsignals undergoes moderate rotation (3°) or scaling (4%). The discussed framework is a fairly robust, single-step method for local depth estimation. We present results for synthetic as well as real image pairs.

Keywords: depth estimation, stereopsis, cepstrum, primary visual cortex, ocular stripe maps

Multiframe analysis of images, such as stereopsis and time-varying image sequences, has been a primary focus of activities within the last decade of computational vision research. In both areas, the key problem has been identified as finding the correct correspondences of pairwise related (i.e. homologous) image points which represent a single point in the physical scene. The so-called correspondence problem has not been solved to-date to apply for general purpose vision tasks. The majority of approaches can be roughly classified into area- and feature-based methods.

- *Area-based techniques* can be identified as methods which apply statistical measures, e.g. various correlation-based similarity measures. These measurements are evaluated to determine corresponding local regions in the two image frames by maximizing the similarity within appropriately selected regions of interest.
- *Feature-based techniques*, on the other hand, have been developed for scenes which contain discontinuities in the multi-parameter function that captures the physical and photometrical aspects of the scene structure sensed. The image features commonly used are contrast edges and their local attributes or higher order measures such as, for example, grey level corners.

For a review of relevant techniques for finding stereo correspondences, see elsewhere[2–4].

In case of stereopsis, the central and – with respect to the general case – not successfully solved problem is that of reliably finding correspondences between two static image frames. To reduce the candidate set of possible correspondences between the two image frames, many authors have proposed use of the epipolar constraint. In the case of two parallel (coplanar) images, the epipolar geometry results in two horizontal lines with corresponding vertical positions: in other words, corresponding points always have the same horizontal positions in the left and right image frame† (this fact is used, for example, by Grimson[5,6] and Marr[7].

*A preliminary version of this article was published in the *Proceedings of the 2nd International Workshop on Robust Computer Vision* (March 8–12 1992), Bonn, Germany[45].

Universität Hamburg, FB Informatik, AB KOGS, Bodenstedtstr. 16, W-2000 Hamburg 50, Germany (Email:ludwig@kogs26.informatik.uni-hamburg.de)

Paper received: 2 September 1992; revised paper received: 5 July 1993

†This is true for the ideal mathematical case. For discrete noisy data, it is necessary to introduce a small horizontal band of tolerance to take into account the distortions in the localizations of corresponding features[6].

0262–8856/94/010016–20

Finding stereo correspondence can be identified as a mathematically ill-posed problem in the sense of Hadamard, which has to be regularized utilizing constraints imposed on the possible solution[8]. The majority of computational approaches are therefore formulated as finding a solution in a high dimensional search or optimization space by minimizing a functional which usually takes into account a data similarity term as well as a model term (e.g. for achieving smoothness) so as to regularize the solution[8,9]. The results of these computational methods – which almost always have been formulated as iterative processes – yield a sparse or dense disparity map.

To avoid the complexity of most of the existing computational techniques we have investigated biological findings about architecture and mechanisms for seeing stereoscopic depth.

The rest of the paper is organized as follows: the next section provides an introduction to the relevant biological material. After that, the mathematical motivation of the method is presented. It is followed by the central part of the paper, which contains the analysis and extensions of the implemented disparity estimation method.

In that section we discuss computational as well as geometrical aspects of the data arrangement in the ocularity stripes. Specifically, we investigate the following:

- noise sensitivity and robustness of disparity estimation;
- the sensitivity of the method for different luminance levels of the two half signals and the inversion of one of the subsignals;
- violation of the mathematical assumption of a pure translational shift between the two subsignals, i.e. the robustness of the cepstral filter when rotating or scaling one of the subsignals; and
- the problems imposed by the assumption of abrupt changes in the ocular dominance profile.

As it is evident from the physical structure of real scenes, the information relevant for local disparity is contributed from the finer details of the surface structure, i.e. signal components in the middle and higher spectral bands of the signal. From this it follows that the method will benefit from an appropriate LoG (Laplacian of Gaussian) bandpass filtering step, eliminating low band background variations and very high band noise components. It has been shown[10] that the effect of the cepstral operator is a squared autocorrelation function with an adaptive prefiltering step. We have investigated the filter further, and our experiments suggest that this prefilter has a bandpass-like effect. To demonstrate the usefulness of our approach, we present results derived for synthetic as well as real camera images.

BIOLOGICAL BACKGROUND

In this section we give a brief overview of the biological material on which we base our computational approach for depth estimation. Since we have tried to decouple technical and biological aspects as far as possible, this section may be skipped by those readers who are interested solely in the technical discussion. This section is not intended to be fully comprehensible without some further background knowledge. It should provide a reasonable trade-off between an introduction to the biological material (for readers unfamiliar with this material) and a refresher (for the others), but it would be beyond the scope of this paper to review all the details. More details can be found in the specialized literature on this subject[11–13].

Computational maps

An alternative to the most commonly realized top-down strategy in computational vision research – which starts from a task analysis (definition of a computational theory), derives an algorithm and ends up with an implementation[7] – is to infer information processing capabilities from the identification of structural principles in the mammalian visual cortex[14,15]. Von Seelen *et al.*[15] introduced the term *neural instruction* set for the identification of a set of structural principles in the organization of cortical areas which are likely to support different processing and behavioural tasks in the living animal or human.

There is evidence for separate mappings of features for different sensory modalities. In the case of vision, several principles of spatial coding, such as patchy retinotopy, columns, stripes and blobs, have been identified (see, for example, Hubel[11], and for a mechanistic interpretation, Mallot *et al.*[14]). However, from the set of maps and principles given above, only the principles of retinotopy and ocular dominance stripe maps are fully established[16].

In addition, a variety of processing streams between dedicated visual areas have bee discovered[12,13]. The specificity of individual channels has been subject of a great amount of psychophysical as well as neurophysiological investigation[12,13,17, 18].

These principles seem to provide a general mapping strategy for different sensory features in terms of coordinate transforms to code features like orientation, colour, ocular dominance, depth or motion in 3D space to positions in a subspace of $\mathcal{R}^2$. These 'computational maps'[12,16] have been postulated to optimally support computational mechanisms of different specificity.

Independent processing streams, ocular stripe maps, and fixation of an environmental point

With respect to the early processing steps from the retina to the primary visual cortex (areas 17 (V1) and 18 (V2), three more or less independent processing streams roughly characterized as the high-resolution form (*parvo-interblob*), low-resolution colour (*parvo-blob*) and colour-blind stereo/motion depth (*magno*) channel have been postulated[12,13]. The above-mentioned structure of an ocular stripe map in the

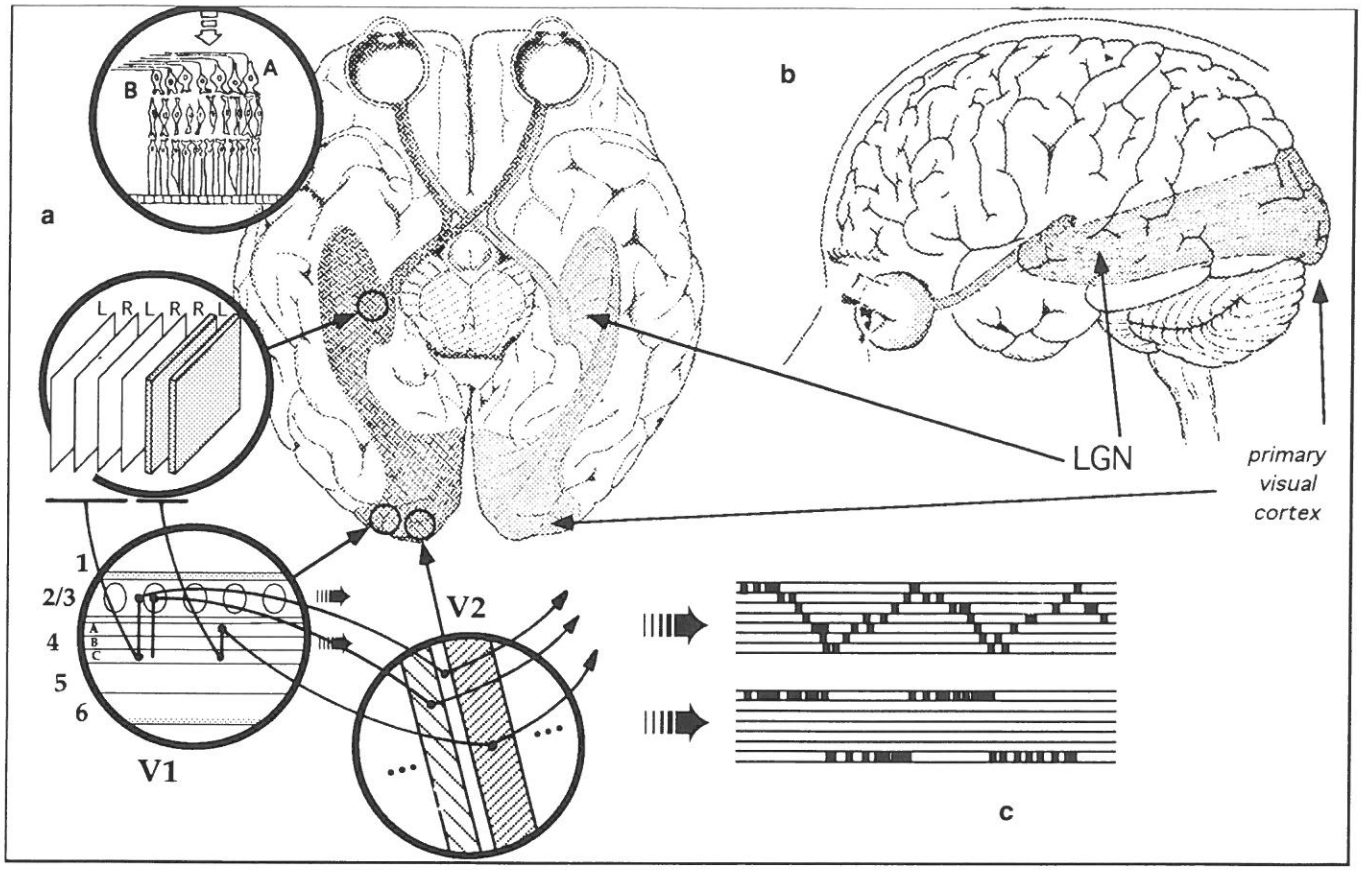


Figure 1 Overview of the general processing streams in the primate visual system (a) Top view of a slice cut through the brain with major visual processing areas (retina, LGN, primary visual cortex areas V1, V2) and segregated processing streams indicated. Letters A and B refer to different ganglion cell types in the retina contributing to the different streams, whereas 4A, 4B and 4C refer to the different regions in layer 4 in V1 (see Livingstone and Hubel[12] for further details); (b) side view of the brain with projections from retina to visual cortex; (c) magnified view of distribution of cells with left/right eye dominance in visual cortical area 17 (V1) in layers 4 (bottom) and 2/3 (top), respectively. As can be clearly seen, the initial sharp tuning of cells in layer 4 (segregated input from left and right eye) is lost in layer 2/3, where we can observe a smooth transition between left and right eye dominance, which indicates binocularity of cells. Image fragments from Livingstone and Hubel[12] and Hubel[11]

primary visual cortex obeys the two-dimensional geometric organization of alternating bands of left and right eye dominance.

Given that these alternating ocular dominance columns are part of the projection path of the magno channel (see Livingstone and Hubel[12], p. 742 and *Figure 1*), and that this magno channel seems to be colour-blind and responsible for the motion and stereo computation in the primate visual system[12], then a key question naturally arises in the context of investigating the neural instruction set of brains. The question about the links between structure and function of a given algorithm or system can be stated in this case as follows: What can a specialized data organization with discrete alternating stripes of left/right ocular dominance of the visual input be useful for?

Figure 2a shows a map based on a preparation of the cortical surface in area 17 of the macaque monkey. To get an impression of the general mapping function and the relevant anatomical locations they have been sketched in *Figure 2b*. The width of these stripes has been measured to be about $W = 0.35 - 0.4$ mm. In early preparations of the monkey striate cortex, an additional substructure, the so-called pale bands, has been made visible. They are of approximately 0.05 mm

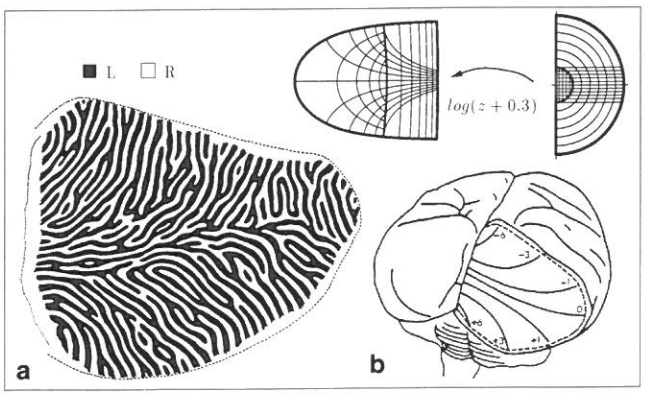


Figure 2 Ocular dominance stripe maps in the macaque monkey. (a) Visualization of alternating bands of left and right ocular dominance columns (view of a flat mounted preparation of a piece of cortex, from Hubel and Freeman[39]; (b) sketch of iso-eccentricity lines mapped to cortical area 17 (corresponding to (a)) and example mapping of retina to visual cortex via a complex logarithmic map (see, for example, Schwartz[44])

width (approx. $W/8$) and separate the ocular dominance stripes[17]. In a later section of this paper we refer to this to give a functional explanation for the existence of these pale bands which, in turn, provide a

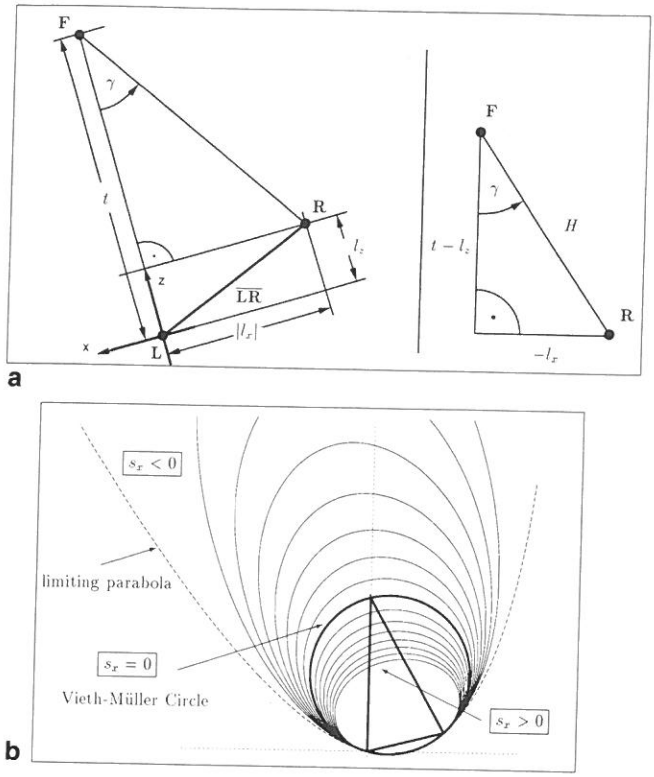


Figure 3 Mathematical description of imaging geometry and iso-disparity lines. (a) Fixating stereo arrangement (L, R, F denote points of left and right eye location and the fixation point, respectively; the 2D reference coordinate system is centred with its origin at L, the left eye location); (b) Lines of constant disparity s_x when using flat retinae (this set of parabolic curves has been determined to serve as a reference set when planar images, such as in cameras, are used instead of ideal spherical images)

compensation for degenerate cases where stereo disparity measurement is erroneous due to occasionally unwanted signal structures.

The usefulness of the peculiar spatial data representation of visual information from the left and right eyes can be explained as follows. Again, with reference to biological vision systems, assume a geometry in which the optical axes of the two image frames fixate a previously identified point in 3D space*.

Then under these conditions of imaging geometry (see *Figure 3* and *6a1*) one can directly construct a circle which is uniquely defined by the two optical centres of projection and the point of fixation. This *Vieth–Müller* circle (sometimes called *horopter*†) defines all points in space that project onto the two retinae with zero disparity (Thales theorem).

Not only projections of space locations with exactly zero disparity can be fused. All 3D points with moderate negative (far field) or positive (near field) disparity within a psychophysically defined region called *Panum's* area (*Figure 6a2*) also contribute to a fused image of varying depth due to the retinal shift of projection to retinal coordinates. Assuming that Panum's area (a fixed fusion range) results from the fixed width of the ocular stripes, as shown in *Figure 6a2*), this area delimits‡ the range of possible local depth estimates which may be computed for a given point of fixation§. 3D spatial locations outside Panum's area produce the well known phenomenon of double images.

Other approaches for local disparity computation have been proposed, e.g. phase-based methods like those presented by Jenkin *et al.*[28,29] and Sanger[30] in the literature. It is an open issue to compare the different approaches in a mathematically rigorous way as well as in their performance on the same image data. Olson and Coombs[10] (pp. 29–30) give some hints how phase correlation is related to the cepstrum filter.

The main points reviewed in this section can be summarized as follows. Besides other channels, a separate channel involved in stereo/motion computation has been identified in the primate visual system. This channel seems to be colour-blind, i.e. conveys only luminance-specific information. The visual information from the left and right eyes is arranged in a salient stripe pattern at a later processing segment of this path.

DISPARITY ESTIMATION

Mathematical preliminaries

The cepstrum of a signal $g(\mathbf{x})$ is the power spectrum of the logarithm of its power spectrum**:

$$\mathcal{C}epstrum\{g(\mathbf{x})\} = \|\mathcal{F}\{\log(\|\mathcal{F}\{g(\mathbf{x})\}\|^2)\}\|^2 \qquad (1)$$

where $\mathcal{F}\{\cdot\}$ denotes the Fourier transform. Two corresponding local image patches extracted from the

*As a part of an active vision system[20,21], a fixating stereo geometry necessarily requires two mechanisms: one attentional control module for the selection of appropriate fixation points; and one for carrying out the vergence movement to fixate (and track) those points. Proposals have been published on how fixation points could be selected – for example, with a computational model[22], using intrinsic or extrinsic features and a channel concept – and how such a point may be tracked in time[23]. Concerning binocular vergence control and real-time stabilization, see also the recent work by Theimer and Mallot[24], and Pahlavan *et al.*[25,26].

†The term 'horopter' is usually reserved for the zero-disparity curve measured for human beings, which slightly deviates from a perfect circle[47].

‡This absolute limit of fusion will be further diminished by the disparity gradient of two neighbouring but disparate points[27], although the underlying mechanisms are currently not clear.

§In the case of idealized circular retinae the iso-disparity lines are circles of different radii, with the Vieth–Müller circle as one element of the set. In the case of flat projection planes, the iso-lines for a fixed disparity s_x are conic sections in the parameters $\mathbf{x}$ and $\mathbf{z}$ following $l_x \cdot \mathbf{x}^2 + (tl_z - l_x^2 - l_z^2) \cdot \mathbf{x} + (l_x + s_x(l_z - t)) \cdot \mathbf{z}^2 + (s_x(tl_z - l_x^2 - l_z^2) - tl_x) \cdot \mathbf{z} + s_x l_x \cdot \mathbf{xz} - 0$ where the loci of zero disparity again form a circle, and the iso-lines for small disparities are elipses. (Here t, l_x and l_z are constants depending upon the particular stereo configuration, see *Figure 3*).

**The cepstrum – an anagram of the word spectrum – is a well-known non-linear filter first used by Bogert *et al.*[32] for the detection of echo arrival times in 1D seismic signals. Due to its simplicity and noise robustness, it has been widely used since then in various application areas from 1D speech processing[33] to solving 2D image registration problems[34]. A comprehensive overview is given elsewhere[35].

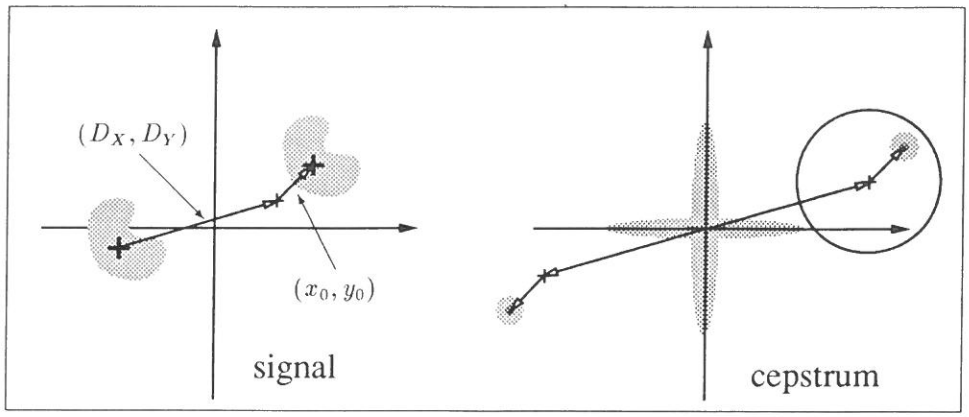


Figure 4 The principle of disparity computation with the cepstral filter. The joint input signal for the Cepstrum is composed of the subsignal $s(x, y)$ and a copy of s which has been transformed by a slight translational shift (x_0, y_0). Since these two subsignals are arranged in a joint pattern (ocular stripe pattern), we can identify a basic shift vector (D_X, D_Y) which denotes the principal offset between left and right dominance column (left). The Cepstrum of the *joint* signal has been shown to consist of the superposition of the Cepstrum of subsignal s and a sum of scaled Dirac pulses which encode the disparity shift. Therefore, one can use the pair of free parameters (D_X, D_Y) to locate zero shift location outside the region dominated by $\mathcal{C}epstrum\{s(x, y)\}$. The vector drawn between this zero shift reference point and the next pulse within a local neighbourhood directly encodes the disparity between the signals from the left and right image (right)

left and right images, respectively, can be arranged in a local neighbourhood to form a single joint signal. This idea has been originally utilized in an algorithm proposed by Yeshurun and Schwartz[1,31] using rectangular patches butted against each other. If such a combined signal is filtered with the cepstrum, the filtered image contains a strong and sharp peak at a position which codes the disparity shift between the two original subsignals. This can be derived mathematically for the ideal case of a pure translational shift. Let $f(x, y) = s(x, y) + s(x - x_0, y - y_0)$; then for the amplitude spectrum of the joint signal it follows that:

$$\begin{aligned}\|F(u, v)\|^2 &= \|S(u, v) \cdot (1 + e^{-i(x_0 u + y_0 v)})\|^2 \\ &= \|S(u, v)\|^2 \cdot 2 \cdot (1 + \cos x_0 u + y_0 v))\end{aligned}$$

holds. Taking the logarithm* yields the following sum:

$$\log(\|F(u, v\|^2) = \log(\|S(u, v\|^2) + \log(2 \cdot (1 + \cos(x_0 u + y_0 v))) \quad (2)$$

This signal is the sum of an image-dependent term (about which no specific judgement can be made in the general case) and an image-independent term (which in principle contains only one dominant frequency component at $\mathbf{w}_0 = x_0 u + y_0 v$. Accordingly, we expect in the power spectrum of $\log(\|F(u, v)\|^2)$ – that is the cepstrum of $f(x, y)$ – a strong point component which results from this second term.

Fourier transformation yields:

$$\mathcal{F}\{\log(\|F(u, v)\|^2)\} = \mathcal{F}\{\log(\|S(u, v)\|^2)\} + \mathcal{F}\{\log(2)\} + \mathcal{F}\{\log(1 + \cos(x_0 u + y_0 v))\} \quad (3)$$

Reminding that the Fourier transform at $\cos(\cdot)$ results in two delta pulses, and since for $\|x\| < 1$ we have the series expansion:

$$\log(1 + x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n}$$

we can transform equation (3) – with $x \equiv \cos(x_0 u + y_0 v)$ – nearly everywhere to:

$$\mathcal{C}epstrum\{f(x, y)\} = \mathcal{C}epstrum\{s(x, y)\} + \sum_{n=-\infty}^{\infty} \frac{\delta(x - n \cdot x_0, y - n \cdot y_0)}{n} \quad (4)$$

Thus we have the following result. The cepstrum of the double signal is the sum of the cepstrum of one of the subsignals and an impulse train with rapidly decreasing amplitude of its peaks. So it should be possible to detect the dominant first peak with a position that codes the shift between the two subsignals[46].

Disparity estimation

Excluding some special cases – which will be discussed later – the portion of the signal corresponding to the first term of equation (4) does not hide the portion of the second term, as depicted in *Figure 4*. Thus the disparity between the two subsignals can be obtained by simple maximum detection in the cepstral plane, as illustrated in *Figure 5*.

When the shift between two subsignals becomes larger, the echo maximum will be more attenuated, since the area of common signal will be increasingly reduced. Lines of constant common signal area define rhombi around the point of zero disparity in the cepstral plane. Reducing the maximum search area to

*'For example, replacing the log step in the cepstral algorithm with a fourth root or arc tangent produces results that do not differ greatly from the standard cepstrum' (Olson and Coombs[10], p. 28).

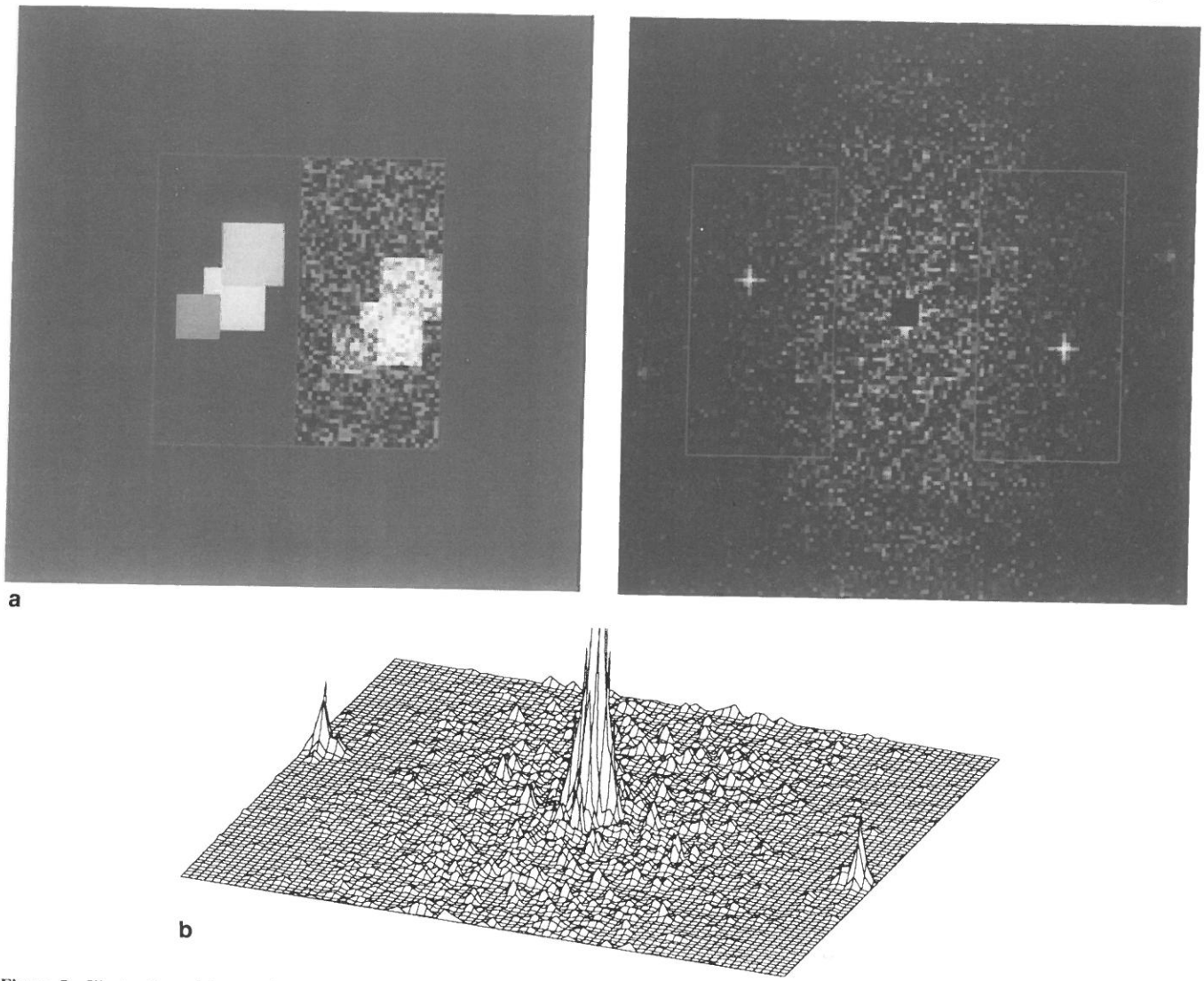


Figure 5 Illustration of the result of cepstral filtering of a joint signal. (a) Joint signal and computed cepstrum. A test pattern taken as the left subsignal has been transformed by a left-downward translation, superimposed by white noise with 100% signal amplitude to generate the right subsignal (left). Cepstrum filter computation for the joint signal $s(x, y)$ displayed as $\log(1+\mathcal{C}epstrum\{s(x, y)\})$ with the central proof being removed (right); (b) three-dimensional plot of $\mathcal{C}epstrum\{s(x, y)\}$. Strong isolated peaks in the left and right half planes – despite the additive noise in the signal – encode the disparity as sketched in *Figure 4*

these rhombi helps to avoid the detection of additional (false) maxima, and reduces the search effort in general by a factor of two or more. As Olson and Coombs[10] pointed out, the number of column transforms required in the last step of the discrete transform can be reduced to a quarter or less.

Furthermore, if the fast Hartley transformation (FHT)[36] is used instead of the usual fast Fourier transform (FFT), the filter output can be computed twice as fast.

Using the cepstral filter method for computing disparities has several nice properties. First, it is fast, because the disparities are computed in a single step without any iterations*. Second, due to the local and therefore independent computation of the disparities, parallelization is easy. Third, it is well-known from previous work that the cepstrum is extremely insensitive to noise[31,35,37], and we will show that the cepstral filter is insensitive to moderate image degradations like rotation or scaling, too.

*It has been shown[10] that this filtering step could be done in 51 ms when using special image processing hardware. With such short computation times it becomes feasible to use the presented cepstrum-based stereo segmentation approach in *active vision* systems for simple obstacle avoidance or object recognition tasks.

ANALYSIS, EVALUATION AND EXTENSIONS

Basic model

The basic model for local depth estimation is based on the subdivision of the cortical plane into stripes with alternating eye dominance (see *Figure 2*). Starting with the idea that a hypothetical disparity sensitive cell uses a local section of two neighbouring stripes as input (see *Figure 6b*) to compute local disparity, we can divide the original left and right images into appropriate

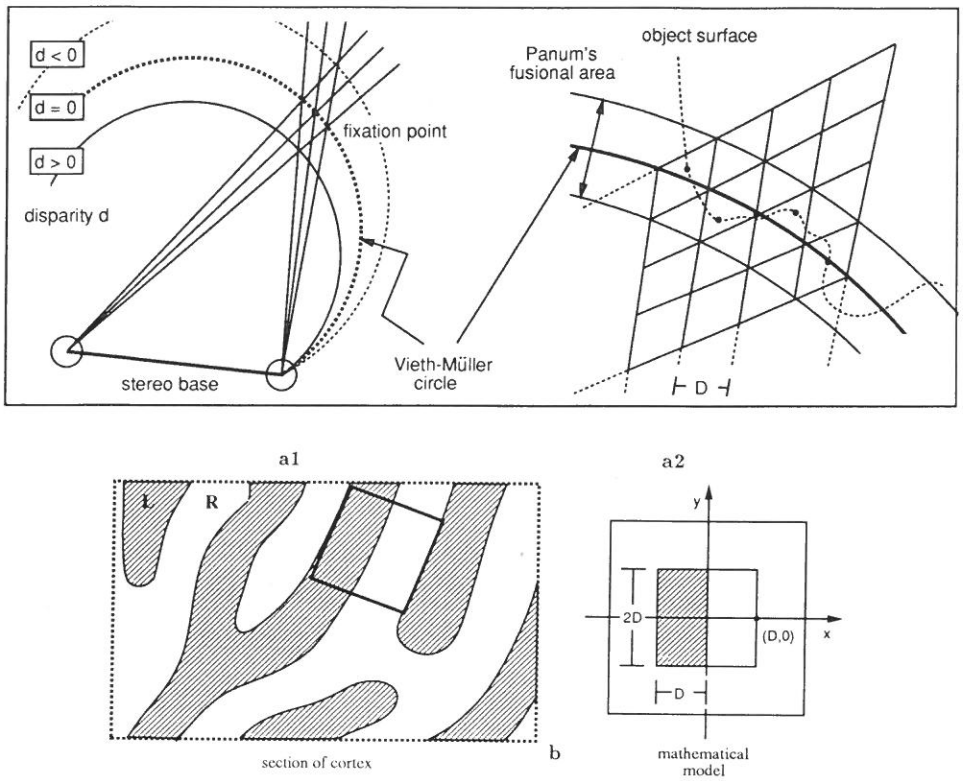


Figure 6 Imaging geometry and modelling of functional components for local disparity estimation. (a1) Fixating stereo arrangement. Spatial locations on the so-called Vieth–Müller circle project to retinal locations with zero disparity. Due to the fixed width of the dominance stripes, disparity can also be computed for 3D points within a small spatial band around the Vieth–Müller circle (see explanation); (a2) fusional area and a hypothetical object surface; (b) a local region of a pair of ocular dominance stripes and the model for disparity estimation. Local disparity is estimated in a square window composed of subsignals each of $D \times 2D$ size (local (x, y)-coordinate system centred in the window)

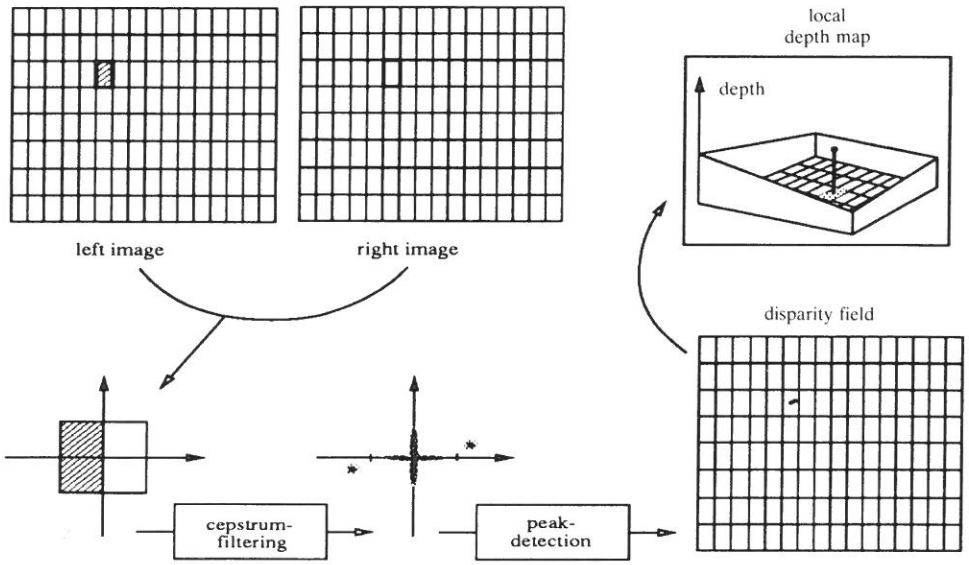


Figure 7 Computational steps for discrete local depth estimation. Referring to the abstract model sketched in *Figure 6*, left and right images are interlaced horizontally to define one image in which the frequency of left/right eye dominance varies with the stripe thickness D. To define localized windows for disparity estimation (based on the cepstrum technique) each pair of bands is divided into regions of $2D$ height

rectangles (see *Figure 7*). Given the disparity at all single locations, we obtain a disparity map from which a (relative) depth map can be easily inferred, as depicted in *Figure 6a2*.

To determine the values of the parameters of the technical model, we have evaluated the relevant and sometimes diverging biological material from various sources[11,38–41] to get a reasonable and consistent parameter setting. It is beyond the scope of this paper to review all the material, data and methods used in detail.

It should be noted here that we collected this data only to get a hint for reasonable values for our purposes (application in a human-like vision system setting), and that the values might be different in other application areas (e.g. for making depth measurements in cases where the image pair has been produced by an electron microscope or the like).

As a result of our investigation, the following parameter setting is used during our computational experiments. The angular extent of the single images is about 200 minutes of arc (twice the extent of the fovea[11]). The images are partitioned into 16 stripes with 32 pixels each. Thus the total resolution of the images is 512×512 pixels. This results in a stereo acuity of 12.5 minutes of arc, which fits well the value

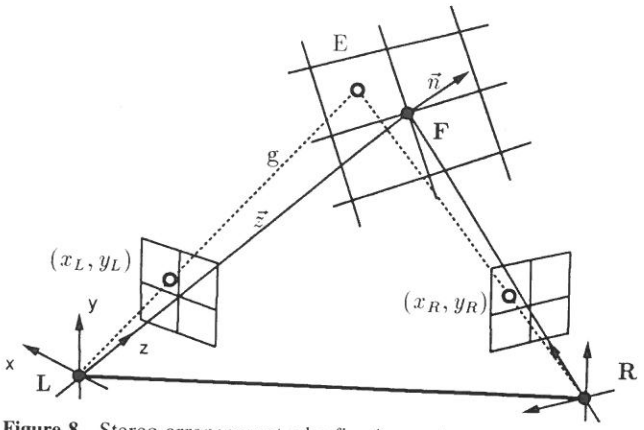


Figure 8 Stereo arrangement whe fixating a plane

of 10 minutes of arc as measured by Tyler[4], and a disparity resolution of 0.5 minutes of arc which is in between the classical values of 0.1 and 1 minutes of arc for hyperacuity or standard acuity, respectively.

Imaging geometry for fixation and distortions in the disparity field

It is reasonable to assume that physical surfaces in the natural environment are smooth, and hence can be approximated locally by their Taylor series expansion. First and second order Taylor series approximations result in planes and second order surfaces, respectively. The following presents the analytical and computational results for fixation of first and second order surface approximation.

Fixation of planes

Consider a stereo arrangement like the one illustrated in *Figure 8*, where the y axes of the left and right coordinate system are aligned. For a given point in 3D space, let the left image coordinates be $\boldsymbol{l} = (x_L, y_L)$. Then the right image coordinates $\boldsymbol{r} = (x_R, y_R)$ can be computed to be:

$$x_R = \frac{a_1 x_L + a_2 y_l}{b_1 x_L + b_2 y_L + b_3} \quad \text{and}$$

$$y_R = \frac{c y_L}{b_1 x_L + b_2 y_L + b_3} \tag{5}$$

where a_i, b_i and c with:

$$a_1 = ft(n_z(t-l_z) - n_x l_x), \quad b_1 = n_x(l_x^2 + l_x^2) - t(n_x l_z + n_z l_x)$$
$$a_2 = -ftn_y l_x, \quad b_2 = n_y(l_x^2 + l_z^2 - l_z t)$$
$$c = fn_z tH, \quad b_3 = fn_z(l_x^2 + l_z^2 + t^2 - 2l_z t)$$

are constants with respect to the given stereo arrangement and local surface orientation (the other constants are explained below). The disparity may be defined then as $\boldsymbol{s} = \boldsymbol{r} - \boldsymbol{l}$.

This result may be derived by the following two major steps: (1) The general formula for the left and right image coordinates of an arbitrary 3D point is derived. (2) The intersection point of the plane E and the line g is substituted in that first formula. Since the derivation of equation (5) is straightforward (but troublesome due to the many constants involved), we present here the relevant substeps, so that the interested reader is able to reconstruct the complete derivation*.

Let us start step (1) with the derivation in the xz-plane ($y = 0$) of the arrangement, with the origin of the left coordinate system at **L** and the other constants be named as given in *Figure 3a*. The point $\mathbf{P}_L = (x, z)$ in the left coordinate system has the left image coordinate:

$$x_L = f \cdot \frac{x}{z} \tag{6}$$

where f denotes the focal length.

The right image coordinate can be computed by the same formula after translation of the coordinate system to $\mathbf{R} = (l_x, l_z)$ and rotation by $-\gamma$ (see *Figure 3*). Translation yields:

$$\mathbf{P}' = (x - l_x, z - l_z) \tag{7}$$

Rotation by $-\gamma$ results in:

$$\mathbf{P}_R = (\cos\gamma(x - l_x) - \sin\gamma(z - l_z), \sin\gamma(x - l_x) + \cos\gamma(z - l_z)) \tag{8}$$

We have for x_R in the right coordinate system (analogous to equation (6):

$$x_R = f \cdot \frac{\cos\gamma \cdot (x - l_x) - \sin\gamma \cdot (z - l_z)}{\sin\gamma \cdot (x - l_x) + \cos\gamma \cdot (z - l_z)} \tag{9}$$

If the third dimension is added now, we get analogously:

$$y_R = f \cdot \frac{y}{\sin\gamma \cdot (x - l_x) + \cos\gamma \cdot (z - l_z)}$$

Substituting the relations $\cos\gamma = (t - l_z)/H$ and $\sin\gamma = -l_x/H$ from *Figure 3a, right*† and simplifying[M] yields:

$$x_R = f \cdot \frac{l_x t + (l_z - t)\, x - l_x z}{(l_x^2 + l_z^2 - l_z t) - l_x x + (t - l_z)\, z}$$
$$y_R = f \cdot \frac{Hy}{(l_x^2 + l_z^2 - l_z t) - l_x x + (t - l_z)\, z} \tag{10}$$

*We have been using a symbolic maths package for troublesome parts of some mathematiocal calculations presented in this paper. We use the symbol [M] (for MATHEMATICA) to indicate that a step might involve a longer manipulation of the equation and cannot be followed directly by every reader. (MATHEMATICA is a registered trademark of Wolfram Research Inc., Champaign, Illinois 61826-6059, USA.).

†Considering in the sequel only geometrical arrangements, where this substitution is valid, of course.

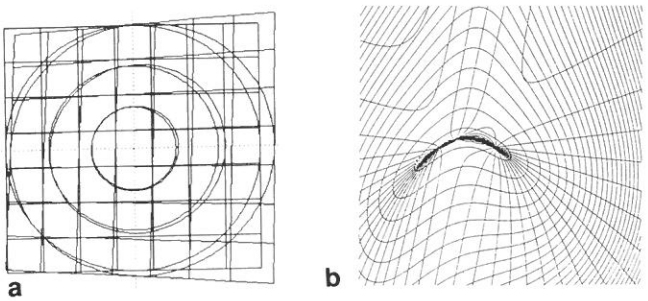


Figure 9 Deformation of a reference pattern after backward transformation from left to right image frame (illustration of formula (5). (a) Distortion of a rectangular grid after transformation from the left to the right retina utilizing a backward projection onto an object surface (the object surface has been approximated by a plane); (b) illustration of the disparity values with isolines. Lines of same absolute disparity value (partially closed curves) and lines of same disparity direction (left and right picture with 100° view angle)

We now turn to the second step of computing the 3D point **P** lying on the fixated plane E given its image coordinates $\boldsymbol{s}_L = (x_L, y_L)$. **P** is just the intersection point of the 3D line g through the origin (0, 0, 0) and (x_L, y_L, f) and plane E. Assume that the fixated surface plane E and the projection ray g of **P** lying in the plane are given by:

$$E: \boldsymbol{n} \cdot \boldsymbol{x} - \boldsymbol{n} \cdot \boldsymbol{z} = 0$$

$$g: \mathbf{0} + \lambda \cdot \boldsymbol{s}_L$$

where $\boldsymbol{n}$ denotes the normal vector of the plane, $\boldsymbol{z}$ is a point in the plane as indicated in *Figure 8*, and λ, $\boldsymbol{x}$ are free variables. Substituting a point from g into E and solving for λ yields a specific λ_s:

$$\boldsymbol{n}\lambda_s\boldsymbol{s}_L - \boldsymbol{n}\boldsymbol{z} = 0 \Rightarrow \lambda_s = \frac{\boldsymbol{n}\boldsymbol{s}}{\boldsymbol{n}\boldsymbol{s}_L} = \frac{n_z t}{n_x x + n_y y + n_z f}$$

Substituting this intersection point $\mathbf{P}_L = \lambda_s \cdot \boldsymbol{s}_L$ in (10) and simplifying[4] the expressions, we arrive at equation (5).

Figure 9a illustrates the effect of this formula, showing how a square grid on the left 'retina' will be distorted when back-projected onto the (planar) surface and then sensed from the position of the right 'retina'. *Figure 9b*, in turn, shows lines of equal absolute value and direction of disparity. As can be seen, the resulting disparity computed by applying the cepstral filter will yield a mean value of all the different disparities lying within the chosen window size. This means that the disparity computation is only applicable near the point of fixation (fovea), where the disparities do not have such a dynamic range so that the mean value in turn becomes erroneous due to the averaging (see *Figure 10*).

Fixation of second order surfaces

The general formula for this case is much more

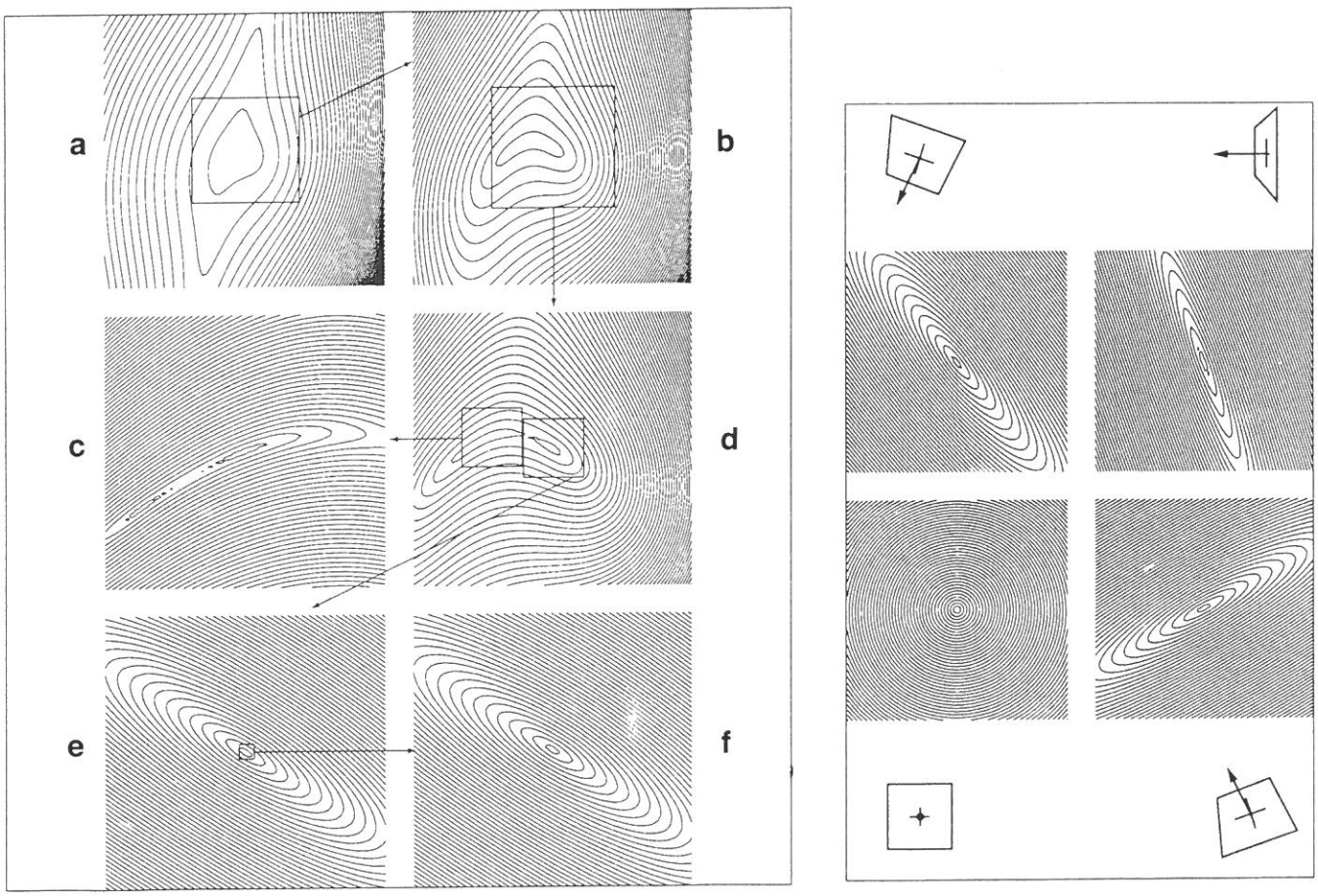


Figure 10 Lines of same absolute disparity value. Left: For different view angles: (a) 160°; (b) 125°; (c) 20°; (d) 80°; (e) 20°; (f) 1°; and right: For different normal vectors of the plane (indicated by small figures) and 1° view angle

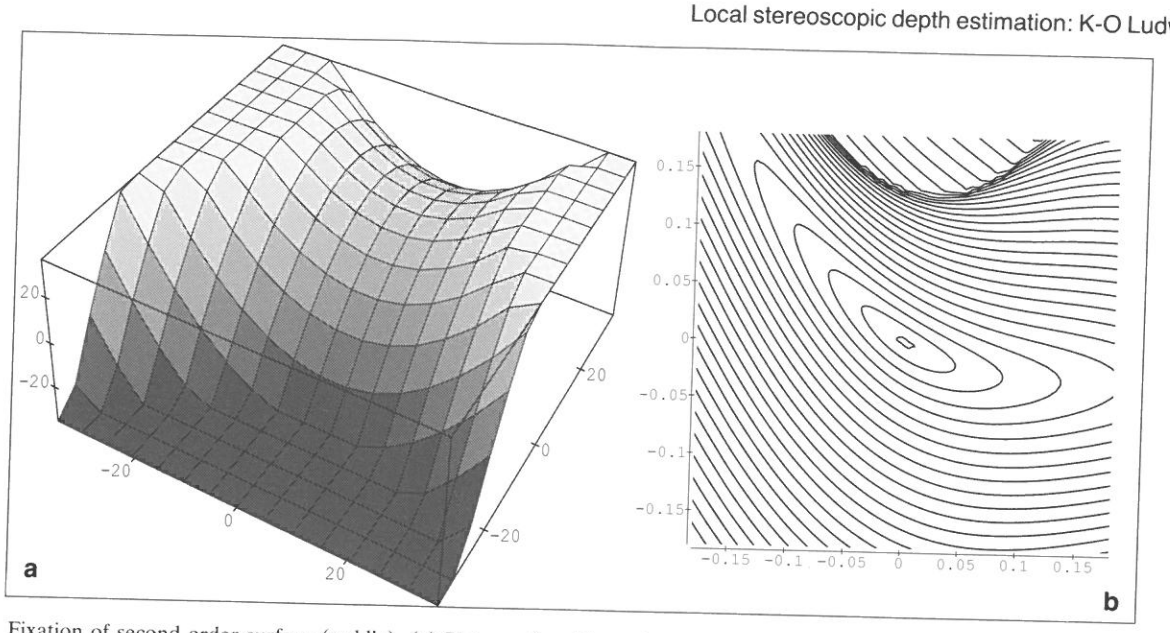


Figure 11 Fixation of second order surface (saddle). (a) Picture of saddle surface. The point of fixation is (0, 0); (b) lines of same absolute disparity value. View angle is 20°. The imaging situation corresponds to that depicted in *Figure 10e* in also using the same surface orientation relative to the view direction. For view angles of about 1°, the distribution of iso-disparity lines for the fixated second order object surface is indistinguishable from those of the planar surface patch (see text)

complicated, since partial occlusion depending upon the particular surface is possible. *Figure 11* shows the iso-disparity lines for fixation of a saddle surface whose partial derivatives have been chosen to be the same as those of the plane in *Figure 10e* for comparison. We found that for foveal angles of about 3° the iso-disparity lines resemble those for the fixation of planes. For 1° they are optically indistinguishable.

The disparities, in general, do not depend upon the particular structure of the surface but primarily on the slant and tilt (i.e. the values of the first order components) of the surface at the fixation point (see *Figure 10, right*). The other important points of the analysis can be summarized as follows:

- The horizontal component of the disparity vector is not dependent upon the height (y-coordinate) of the underlying 3D point. The vertical component depends upon all three coordinates of the surface point, although in the neighbourhood of the fixation point the horizontal component of the disparity vector clearly dominates.
- The local planar approximation of the fixated surface results in a simple function for the disparity shift (equation (5). The iso-disparity maps show that in the case of a foveal analysis, the only point with zero disparity is the point of fixation, whereas this is not true for non-foveal imaging situations.
- The variation of the disparity magnitude can only be neglected within a very small view angle, in order to be reliably estimated with ocularity stripes of constant width. Therefore, the size of the patches which contain signal components from the left and right eye, must not be chosen too big, since otherwise (mean) disparity values become useless due to a significantly diverging local field of disparity values within the patch.

Identification of limits and specific problem cases of the cepstral filter

In this section we give an overview and short discussion of the problems the disparity estimation technique has to deal with. We tested the method with respect to several topics. The results have served as the starting point for a major investigation towards improvement and extension, as described in the next section.

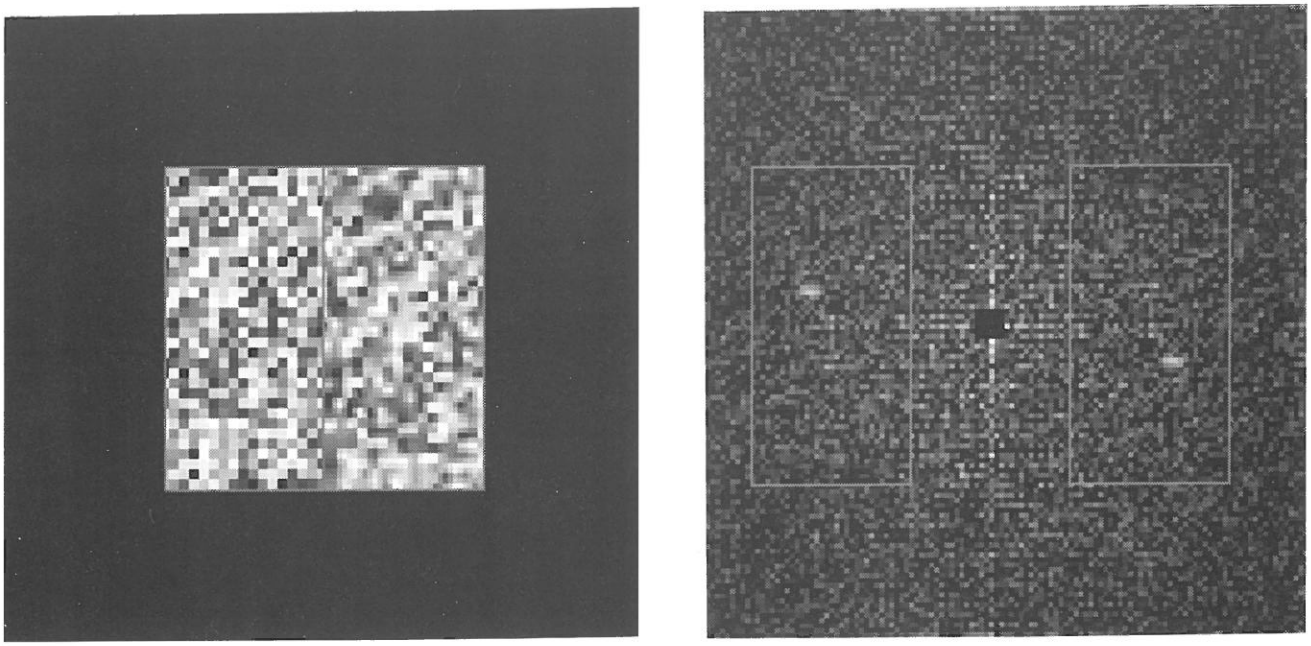
Noise robustness

The cepstral filter has shown to be extremely robust against noise[31,35], although the value of the signal-to-noise ratio (SNR) alone is an insufficient measure of performance[35]. Further mathematical analysis has shown that the relative bandwidth of the signal and the noise can be also important[42]. Despite this fact, a general rule of thumb can be given for real image data. The correct peak of the filter output can be detected with additional white noise up to 100% of the signal amplitude.

Robustness against signal deformations

For the general case of perspective projection the assumption of a pure translational shift in the transform between the two subsignals extracted from left and right images is violated. It has to be investigated how the method behaves if one of the subsignals is distorted. To compare the results with values given for the human visual system, and to have an easy parameterization for distortions, we decided to investigate rotations and size changes, and combinations of both effects.

Our evaluation, described in this subsection, has shown that the disparity estimation with the cepstrum remains correct under rotation or scaling of one of the



a b

Figure 12 Filtering of a disparate signal with additional rotation of right subsignal. (a) A double signal composed of a left part (randomly generated) and a right part rotated by 4° and shifted by a fixed amount; (b) discrete cepstrum of (a). The cepstral peak is weaker and smeared around the true value

subsignals as long as the values do not exceed 3° or 4%, respectively. These two figures have been derived using computer generated random dot stereograms (RDS), since this type of signal possesses a single sharp peak for disparity detection which is located in exactly one pixel of the cepstral plane. Using other than RDS signals might have biased the evaluation, since the particular form of the cepstral peak would have influenced the disparity estimation in one way or the other.

We have chosen every random square to be a 2×2 pixel. This is small enough not to make the random squares a significant structure themselves, and big enough to have a fair representation of the rotated and scaled version. The size of the left subsignal is 32×64 pixels, thus containing 512 random squares. Based upon the values given by Yeshurun and Schwartz and the results of our own test runs, we decided to investigate the performance for rotations $\varphi \in [0°, 8°]$ and scaling factors $s \in [1.0, 1.15]$ as well as combinations of both using a discretization of $\Delta\varphi = 0.5$ and $\Delta s = 0.01$, respectively*.

At each point (φ, s) 50 experiments have been performed for the evaluation to get a reliable statistical basis. In every experiment the left part of the double signal has been generated randomly, whereas the right part has been a rotated and scaled version of it. The right subsignal has additionally been shifted by an arbitrary but known fixed amount (e.g. $(s_x, s_y) = (5, 7)$). The individual pixels of the rotated or scaled subsignal (see *Figure 12* as an example) have been computed from the left subsignal by 4×4 oversampling and averaging of the 16 values.

*A scaling factor of $s = 1.04$ here means that the left signal has been enlarged by 4%.

The evaluation at a point (φ, s) comprises the following values†:

- Since the shift of the right subsignal is known, we could count the number of exact disparity estimates to determine the area in the (φ, s)-plane where the cepstrum technique operates without error.
- The number of disparity estimates located in a tolerance region of ± 5 pixels around the true value has been counted. For these estimates, mean and variance were computed to get information about the specific nature of wrong estimates (see *Figure 14* in comparison to *Figures 13 and 15*).
- For comparison, a smoothed version‡ of the cepstrum was analysed in every single experiment in the same way.

The evaluation procedure described so far has been performed for different implementations of a disparity estimation procedure based on equation (1). One way to compute the disparity (due to Lee *et al.*[43]) is to compute the second term (the impulse train) in equation (4). This method needs three discrete Fourier transforms and some special cases (e.g. division by

†Due to the constant discrete sampling of the value ranges for rotation angles and scaling factors, we can define a 2D discrete lattice for the product of the two sets. For convenience, this data representation is called (φ, s)-plane in the sequel.

‡Two averaging steps of the form:

$$I_{xy} = \frac{1}{9}\left(8I_{xy} + \sum_{\substack{|x-i|=1 \\ \vee |y-j|=1}} I_{ij}\right)$$

were used. The indices xy and ij denote discrete locations in the cepstral plane.

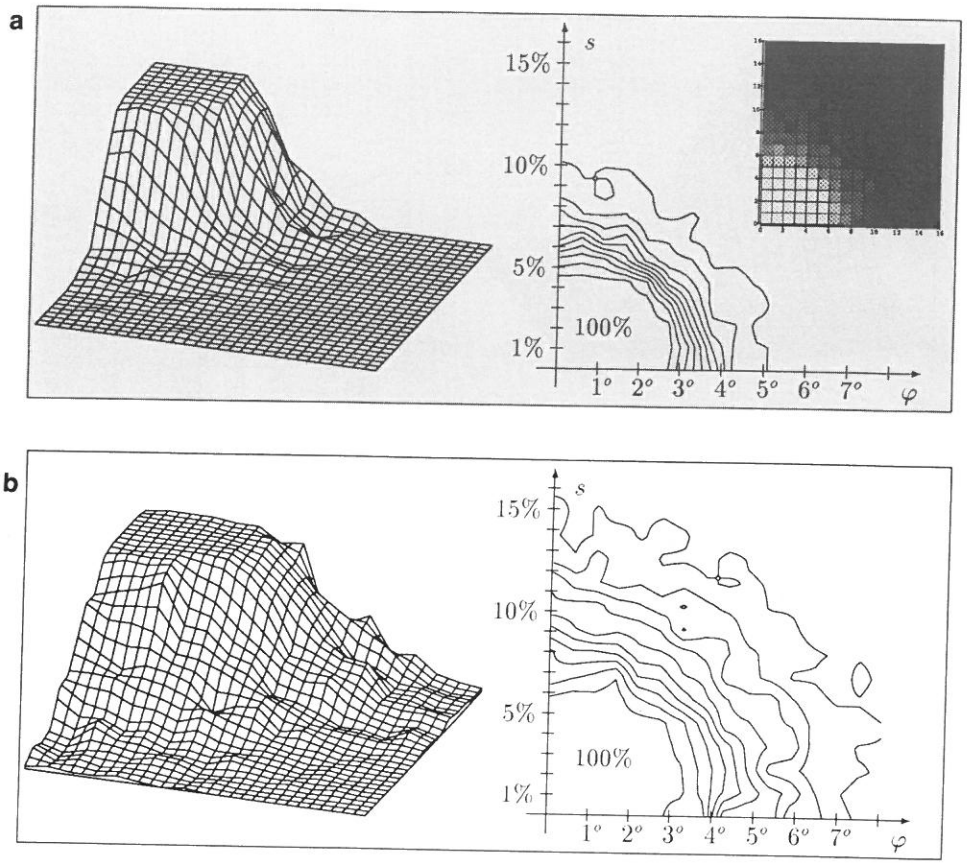


Figure 13 Evaluation of rotation and scaling effects in the disparity estimation using the cepstrum
(a) Percentage of exact disparity estimates. Three dimensional plot (the grid used for visualization is twice the resolution of the data) (left); the same data set as a contour plot (lines of constant percentage in steps of 10%) (right). As can be seen from the contour representation (right), the disparity estimation with the cepstrum yields exact values in a well defined region. Outside this circumscribed area the estimates rapidly become unreliable; (b) percentage of exact disparity estimates using a smoothed version of the cepstrum (see text). Three dimensional plot (left); contour plot (right)

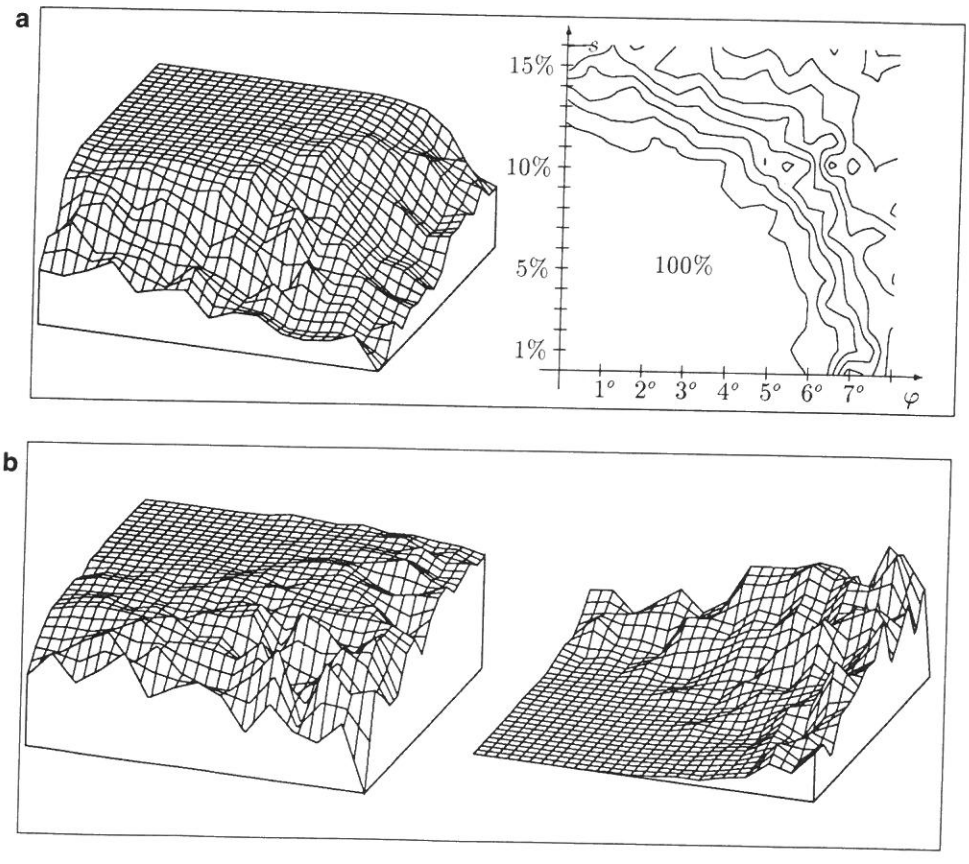


Figure 14 Evaluation of rotation and scaling effects in the disparity estimation using the cepstrum.
(a) Percentage of disparity estimates within a tolerance region of ±5 Pixel (see text). Three dimensional plot (left); contour plot (right);
(b) representations of mean μ and variance σ^2 of the disparity s_x within the tolerance region of ±5 pixels located off the true value (see text) for a smoothed version of the cepstrum. Left: three dimensional plot of $\mu(\mu \in [0, 5.5])$; right: three dimensional plot of $\sigma^2(\sigma^2 \in [0, 13.5])$. As can be expected from the variance data, the wrong estimates have not been completely wrong but only shifted one or two pixels in the neighbourhood. This was confirmed by spot-checks. The data for s_y do not significantly differ from that of s_x

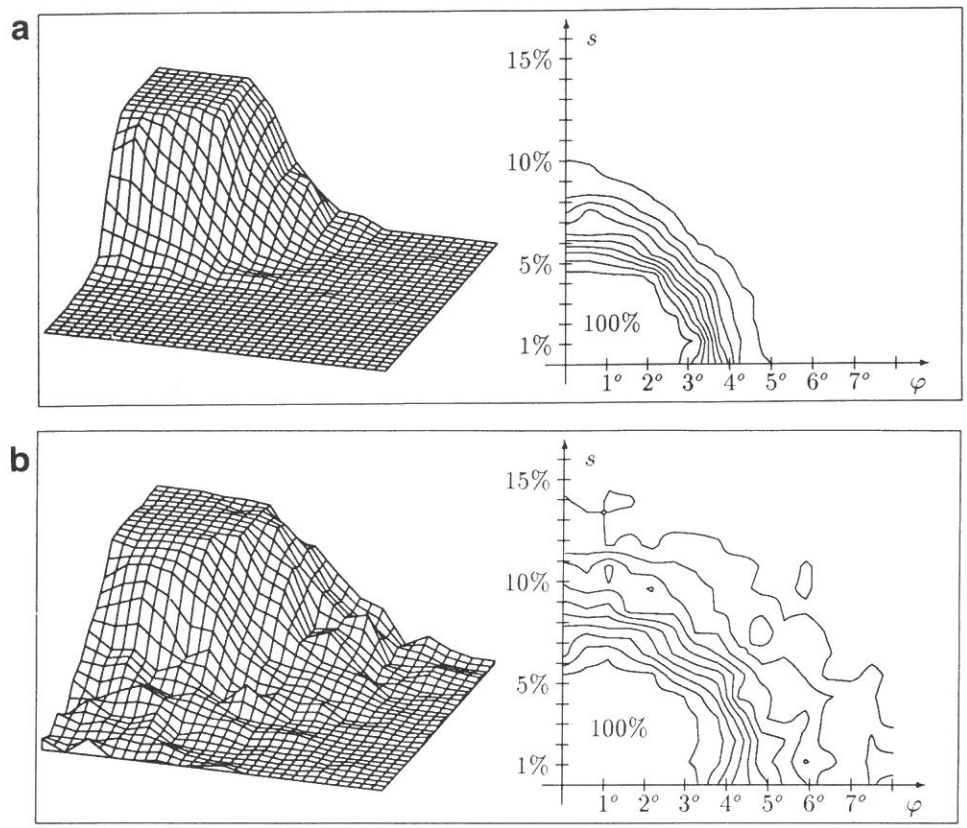


Figure 15 Evaluation of rotation and scaling effects in the disparity estimation using the impulse train method (see text). (a) Percentage of exact disparity estimates. Three dimensional plot (left); same data as a contour plot (right); (b) percentage of exact disparity estimates when using a smoothed cepstrum (see text). Three dimensional plot (left); contour plot (right.) Compare with *Figure 13*

'near' zero values) have to be treated separately*. Yeshurun and Schwartz stated that the peak in the cepstrum is remarkably dominant with no other competing areas of high intensity. This makes it possible to implement the cepstrum filter directly as described by formula (1). Then, only two discrete Fourier transforms are needed.

Figure 13a shows that the disparity estimation is exact for rotations less than approximately 3° and scaling factors less than 4%. If both effects are combined, lines of constant percentage of exact disparity estimates follow the circle $\sqrt{\varphi^2 + s^2} = constant$. It can be further observed that the error free (φ, s)-area of the method is sharply bounded. As can be seen from *Figure 13b* the area of correct disparity estimates can be enlarged when using a smoothed cepstrum at the cost of a more smooth transition between the areas of correct and incorrect disparity estimates. For comparison, *Figures 15a, b* show the same data when using the impulse train method. As can be seen also by comparing the illustrations, the results do not differ significantly.

We found empirically for the case of real camera images that these limits might be extended to 6° and 6%, respectively. Yeshurun and Schwartz[31] (p. 763) reported that the algorithm routinely accepts rotation of 10° and size changes of up to 15%, as a result of their analysis, but these values seem to be about a factor 2 too high for the typical case of noisy images.

The cepstral filter as proposed in the literature with rectangular windowing functions[12,34] has some specific problems, discussed in the following subsections.

*We will refer to his method later in the text, and in captions to illustrations as the *impulse train method*.

Straight edge segments

Besides the well-known *aperture problem* which every local disparity estimator has to deal with, the original algorithm has some problems with linear edge segments, since additional maxima may appear. To understand this phenomenon of additional maxima occurrence we may model the double signal (denoted by $g(x, y)$) for the idealized case as the sum of two rotated rectangular boxes, both degenerated to pulses in the direction orthogonal to the elongation direction (see *Figure 16*).

The Fourier transform of $g(x, y)$, by utilizing the scaling and shift theorem, can be computed to be (see *Figure 16* for an explanation of the constants a, b, x_0, y_0):

$$G(u, v) = \frac{1 - e^{-iau}}{2\pi iu} + \frac{1 - e^{-ibu}}{2\pi iu} \cdot e^{-i(x_0 u + y_0 v)} \qquad (11)$$

and the power spectrum after some simplification[M] becomes:

$$\|G(u, v)\|^2 = \frac{1}{2\pi^2 u^2} \cdot [2 - \cos(au) - \cos(bu) - \cos((a - x_0) - y_0 v) - \cos((b + x_0)u - y_0 v) + \cos((b + x_0 - a)\, u - y_0 v) + \cos(x_0 u - y_0 v)] \qquad (12)$$

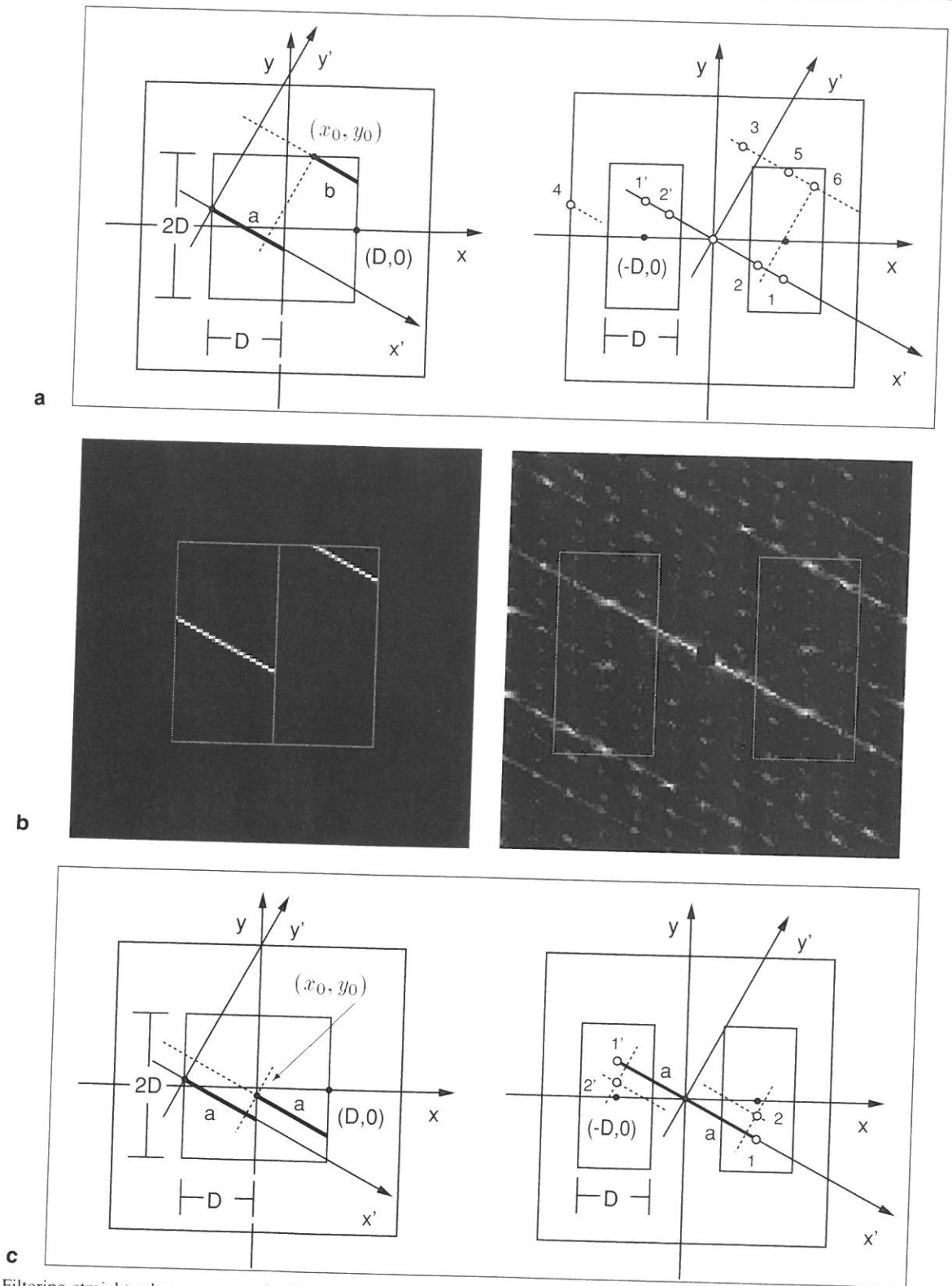


Figure 16 Filtering straight edge segments. (a) Illustration of the local coordinate systems and the parameters used (left). The reference coordinate system has been rotated to get edges a and b (of different length) collinear with the abscissa (x'-axis). The corresponding maxima in the cepstral plane are predicted by formula (12) (right). The numbers 1 . . . 6 denote the corresponding *cos*-component in the formula (numbers with apostrophe simply denote corresponding mirror images); (b) Left: Discrete test signal generated in analogy to (a). Right: Computed cepstrum; (b) in the case of two edge segments of equal length ($b = a$) there are only two additional maxima in the cepstral plane (see formula (13))

We do not need to transform formula (12) further, since a component of the form $\cos(\alpha u + \beta v)$ in the power spectrum corresponds to a peak in the cepstrum at (α, β).

We can now distinguish two cases here, depending upon the values of a and b: it turns out for the general case ($b \neq a$) that up to four additional local maxima may appear within the search area (with dimensions as

proposed by Yeshurun and Schwartz[7]) when the subsignals contain a single straight edge segment (see equation (12) and *Figures 16a, b* for an illustration of the situation).

In the special case ($b = a$), which covers the majority of practical cases, only two maxima appear since the formula then reduces[4f] to:

$$\| G(u, v) \|^2 = \frac{1}{\pi^2 u^2} \cdot (1 - \cos(au))(1 + \cos(x_0 u - y_0 v)) \qquad (13)$$

See *Figure 16c* for illustration.

Whether these maxima are strong enough to disturb the disparity estimation depends upon the search area used and in how far the idealized model fits the concrete line structure in the raw image.

Varying illumination and signal inversion

Another question is how varying lighting conditions are tolerated. Yeshurun and Schwartz[1] (p. 763) state that:

> "The algorithm [. . .] was not disturbed by this intensity difference, nor by simple additive intensity increments of 50 percent to one image of a stereo pair. In fact, positive and negative stereo pairs can be processed with no difficulty, as is evident from the mathematical structure of the cepstral filter."

This line of argumentation is mathematically correct, but does not take into account the practical implementation of the algorithm. We found that different illumination levels (or signal inversion) in either subsignal may disturb the detection of small disparities, when using raw grey level images, since an additional maximum may show up at a position where zero disparities are coded. To explain this additional maximum, consider for a moment only the additive brightness alone without the modulation due to the signals, that is a box shaped signal – for a rectangular pulse in 1D. The spectrum of a rectangular pulse (of width a) is a sinc-function and the cepstrum therefore will have a peak at the positions a and $-a$.

Improvements and extensions

This section contains a description of the major contributions for improving and extending the disparity estimation technique. The topics have been motivated by the results of our analysis in the previous sections.

Varying illumination

Employing the technical counterpart of an observation by Braitenberg[38] made for biological systems*, the reference for zero disparity may be moved away from $(D, 0)$ (see *Figure 6*) to another place in the cepstral plane to avoid the discussed additional maximum due to different illumination or inversion at $(D, 0)$.

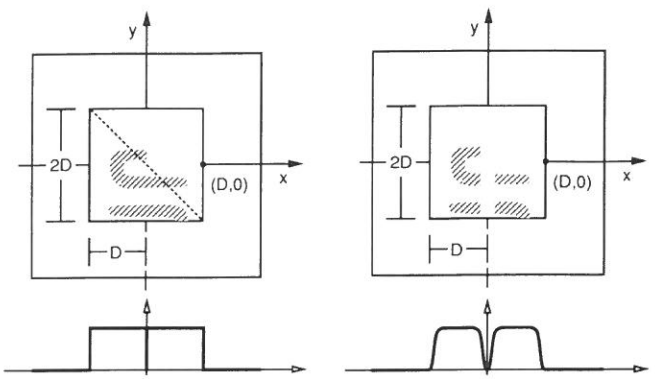


Figure 17 Functional interpretation of *pale bands* as a separating band attenuating the signal from left and right image at the border of the stripes to improve the cepstral filter and to avoid problems in cases with 'artificial structures'. The width of the pale bands has been chosen to be around 1/7–1/8 of the width of the stripes according to the biological data. Left: Conventional support functions (rectangular windows). The dotted diagonal line has been introduced to show that left and right ocular subsignals could have bee composed in arbitrary ways, e.g. as a two-triangles composition. Right: Smoothed windows which separate the two subsignals. At the bottom, one-dimensional profiles taken along the x-axis are shown to demonstrate the difference in the spatial weighting for both window functions

Functional interpretation of the pale bands

Under special unfortunate conditions (see, for example, *Figure 17*) it is possible that 'artificial structures' appear when rectangular support functions for signal extraction are used.

This may disturb the disparity computation occasionally. If the density reduction of horizontal fibres – as observed in the preparation of LeVay *et al.*[19] – is interpreted as a reduction in signal strength from either eye in the neighbouring stripes, the corresponding rectangular support functions in the technical model may be modified accordingly. It turns out that in many cases the problems with straight edge segments and artificial structures can be reduced by modifying the windows, as depicted in *Figure 17*†. This idea can be extended to further improve the cepstral filter.

Other support functions for windowing

The success of the introduction of a separating band in the joint signal can be understood by the following argument: If rectangular support functions are butted against each other, information about the shape of the original subsignals is lost in the joint signal.

Using the separating pale bands can help to preserve this information. With this in mind, we can extend the cepstrum technique further by using other than rectangular support functions to improve the signal properties. In addition, it is useful to weight the disparities, since – ideally – we want to determine the

*". . . it seems that the cuts in one picture are halfway between the cuts in the other picture, so that each strip has overlapping information with the stripes on either side, belonging to the other eye . . ." Braitenberg[38] (p. 386).

†We used appropriate values according to the measurements given in Le Vay *et al.*[19].

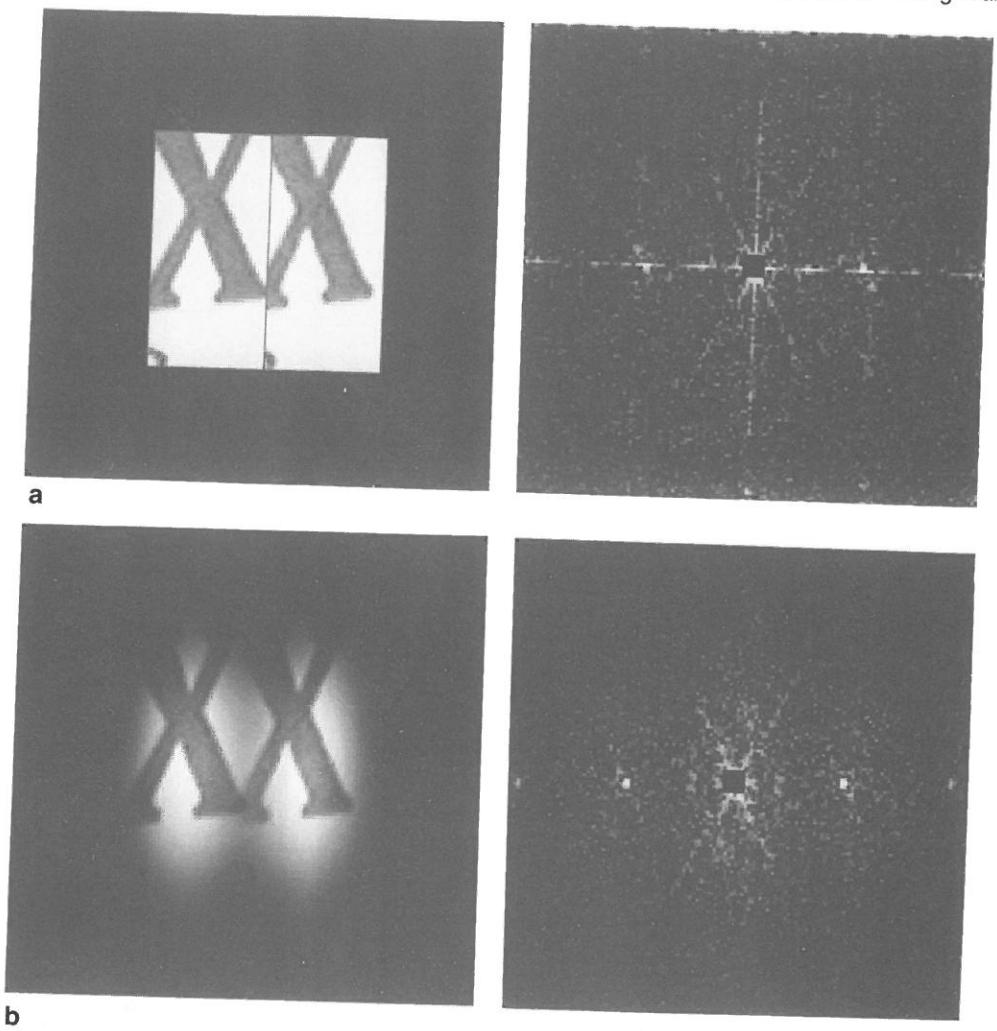


Figure 18 Gaussian window functions improve the cepstral output. (a) Rectangular window functions; (b) Gaussian window functions. As can be clearly seen, artifacts due to the bad signal properties of the rectangular windows are eliminated when using more smooth windows like Gaussians. The second maximum term of the impulse train outside the search area (see equation (4)) also becomes visible. The Gaussian window functions have been generated using $\sigma_x = 10⅔$ and $\sigma_y = 21⅓$

disparity at one – the central point – of the window, using the neighbourhood only for structural/textural support.

We demonstrate in this paper that the use of Gaussian windows – which fulfill both conditions stated above – for the extraction of the local left and right image information produces a more easily identifiable maximum (see *Figure 18b*).

Furthermore, the use of different support functions for the extraction of the left and right subsignals, respectively – as an approximation for a circular receptive field of a disparity sensitive cell is also feasible (see *Figure 19a*). The results are also better than those with standard rectangular support, and the additional effects due to straight edge segments or varying illumination are greatly reduced.

LoG filtering considered beneficial

The cepstrum technique is compatible with any preprocessing operation[1]. LoG filtering is often considered as a technical approximation to the centre-surround architecture of retinal ganglion cells[7]. Since LoG filtering corresponds to computing a derivative of the smoothed signal, one can hope to sharpen the cepstral peak when using an appropriately small σ, where σ is the standard deviation of the Gaussian. We found in particular that a σ of about 0.35 to 0.71 improves the results further within the standard setting.

The cepstrum and autocorrelation

Based on the following formula due to Olson and Coombs[10] (p. 28):

$$\log(\|F(u,v)\|^2) = \left\| \frac{\sqrt{\log(\|F(u,v)\|^2)}}{\|F(u,v)\|} F(u,v) \right\|^2 = \|H\|^2 \tag{14}$$

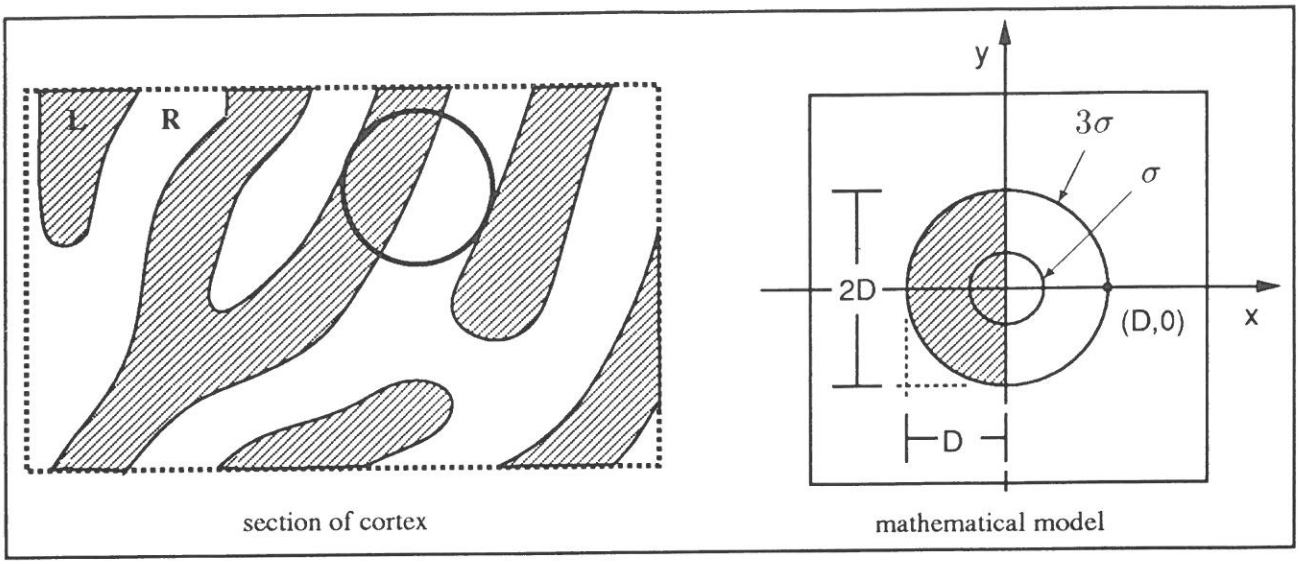
and further:

$$\delta\mathcal{F}\{\|H\|^2\}\|^2 = \|\mathcal{F}\{H^* \cdot H\}\|^2 = \|h^* \star h\|^2 = \|h \circ h\|^2 \tag{15}$$

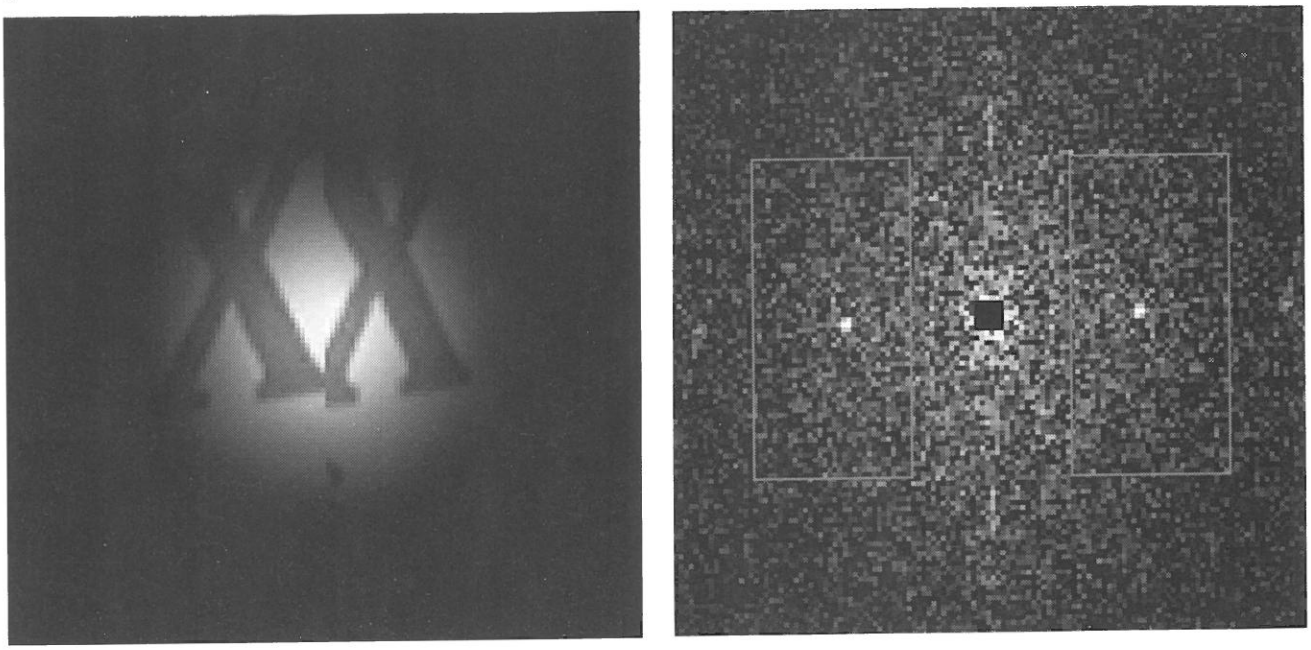
The cepstrum of $f(x, y)$ can be written as:

$$\mathcal{C}epstrum\{f\} = \|h \circ h\|^2 \tag{16}$$

with $h(x, y) = k_F(x, y) \star f(x, y)$, where $\star$ and $\circ$ denote the convolution and the correlation operator, respectively, and h^* is the complex conjugate of h.



a



b

Figure 19 The use of different Gaussian support functions in the left and right image is also feasible, and yields better results than standard rectangular windows. (a) Sketch of area 17 with a circular receptive field of a hypothetical cell for disparity computation; (b) cepstrum with different but complementing Gaussian window functions. The Gaussian half windows have been generated using $\sigma = 21\frac{1}{3}$

As can be seen from this equation, the cepstral filter can be regarded as an autocorrelation operation preceded by a non-linear and image-dependent filtering step[10]. It is interesting to know how the prefiltered image h and the kernel k_F look. For many examples the prefiltered image looks approximately like a bandpass filtered version of $f(x, y)$ with a narrow kernel. We compared the cepstrum results with autocorrelation applied to appropriately LoG filtered images. We found the performance of the cepstral filter substantially better than the autocorrelation results.

Experiments with synthetic images and real world image data

We generated image pairs with a software package for computer graphics visualization, which utilizes a ray projection technique (ray tracing) to investigate the precision of the cepstral disparity estimates under precisely known conditions. It turns out that the mathematical values given by equation (5) can be computed exactly in most cases. Slight deviations* – depending upon the texture of the plane – in about 10% or less of the disparity estimates for a given image pair are due to two effects. First, due to the discrete sampling of the textured surface by the raytracer and

*These deviations have been at most one pixel difference from the correct values when using 512×512 images for the evaluation. The accuracy in depth is dependent on the particular stereo arrangement at hand and can be computed from this value and the distance of the fixated target.

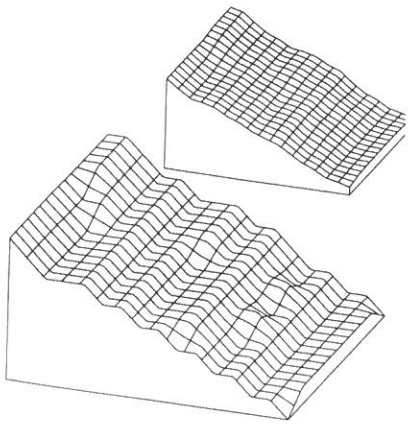
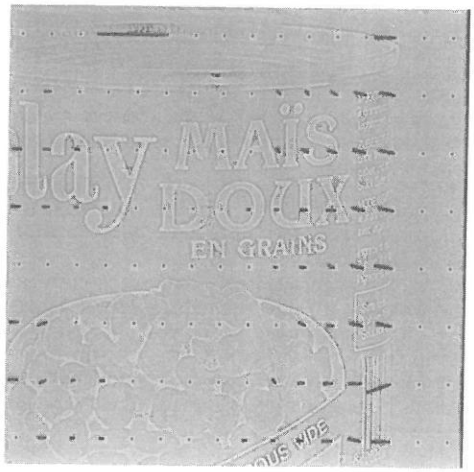


Figure 20 Disparity field (left) and local depth map (right) for the camera image 'book page'. Computation with equal Gaussian window functions for left and right subsignals of each pair of dominance stripes ($\sigma_x = 10⅔$, $\sigma_y = 21⅓$; the rectangles only outline the subdivision of the image). The depth map is shown for the raw data (surface plot, bottom) and as a smoothed version utilizing a 4-point adjacency in a 3×3 neighbourhood (surface plot, top)



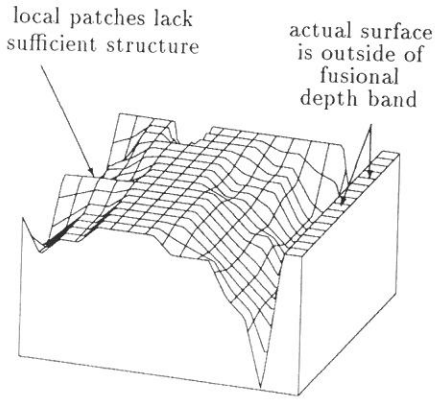


Figure 21 Disparity field (left) and local depth map (right) for the camera image 'maize tin' computed by the improved algorithm using equal Gaussian window functions for left and right subsignals of a pair of dominance columns ($\sigma_x = \sigma_y = 10⅔$; with a previous LoG filtering step with $\sigma = 0.71$). Locations on the surface plot of the depth map have been marked by arrows with the respective condition indicated when the algorithm fails. Both image pairs have been taken at a distance of 2 m with stereo base length 7.00 cm using a precision adjusting device to produce the fixating arrangement. The (foveal) angle of extent is 200 minutes of arc

second, due to the averaging effect caused by the window size of the two parts of the joint signal.

The local depth maps computed by the improved algorithm (with Gaussian support functions) for two real camera images are shown in *Figures 20* and *21*. The image pairs have been taken at a distance of 2 m with stereo base length 7.00 cm using a precision adjusting device to produce a fixating arrangement. The (foveal) angle of extent is 200 minutes of arc (or approximately 3°) in both cases. As can be seen, the qualitative shape of the fixated objects is recovered nicely, despite some local failures of the disparity computation (e.g. due to areas with no texture).

CONCLUSIONS AND PROSPECTS

We have presented a method for how a fixating binocular observer can recover local depth information with a single step computation avoiding the correspondence problem. This method is motivated by recent findings about the architecture of biological visual systems. In contrast to standard formulations of the stereopsis problem, this technique needs neither regularization nor iterative computations to obtain the solution.

Despite its capabilities, the algorithm naturally has its limits, some of which can also be found in other approaches. The presented method has a maximum limit for disparities. The disparities are computed as discrete values derived from the simple peak detection (non-maximum suppression) in the cepstral plane. The algorithm obviously fails if the subsignals do not contain enough structure to contribute to the peak in the cepstrum of the double signal. We also discussed some of the problems of the original cepstrum-based disparity estimator with straight edge segments, different illumination and 'artificial structures'. These limitations, however, have been overcome by our methodological improvements. The algorithm will yield two peaks at depth boundaries, and it depends upon the post-processing steps whether this is acceptable for a given application.

For the technical aspects of the method, a mathematical analysis of support functions with good signal properties is necessary, since we currently investigated only box and Gaussian shapes. In relation to this analysis, the bias introduced from the window shape as an error component in the estimation has to be evaluated. It would also be interesting to investigate in greater detail the prefilter properties of the cepstrum filter.

For the incorporation of this local depth estimation technique in an active vision system, the problem of combining multiple depth maps has to be analysed. With respect to this application, we are investigating the computational efficiency of discrete frequency transforms. We developed a 2D Hartley transform, which is two times faster than ordinary FFT algorithms, on the basis of the 1D Hartley transform as presented by Bracewell[36] to speed up the cepstrum filter, and we are investigating for further improvements.

ACKNOWLEDGEMENTS

The first author wants to thank J-O Eklundh for helpful and interesting comments after the demonstration of an early version of the program. We want to thank also the members of the vision group here in Hamburg for giving comments on the work. The detailed and valuable critique of one of our anonymous reviewers is as gratefully acknowledged, as is K Rohr's proof reading of the final version of the paper. The authors are, of course, responsible for any remaining typos and errors.

REFERENCES

1 Yeshurun, Y and Schwartz E L 'Neural maps as data structures: Fast segmentation of binocular images', in E L Schwarts (ed.), *Computational Neuroscience*, Chap. 20, MIT Press, Cambridge, MA (1990) pp 256–266
2 Baker, H H 'Stereo vision systems', *Proc. Int. Conf. on Cybernetics and Society*, Seattle, WA (October 28–30 1982) pp 322–326
3 Barnard S T and Fischler, M A 'Computational stereo', *Comput. Surv.*, Vol 14 No 4 (1982) pp 553–572
4 Dhond, U R and Aggarwal, J K 'Structure from stereo – A review', *IEEE Trans. Syst., Man. & Cybern.*, Vol 19 No 6 (1989) pp 1489–1510
5 Grimson, W E L *From Images to Surfaces: A Computational Study of the Human Visual System*, The MIT Press, Cambridge, MA (1981)
6 Grimson, W E L 'Computational experiments with a feature based stereo algorithm', *IEEE Trans. PAMI*, Vol 7 No 1 (1985) pp 17–34
7 Marr, D *Vision*, W H Freeman and Co, San Francisco, CA (1982)
8 Poggio, T, Torre, V and Koch, C 'Computational vision and regularization theory', *Nature*, Vol 317 (1985) pp 315
9 Barnard, S T and Fischler, M A 'Computational and biological models of stereo vision', *Proc. Image Understand. Workshop*, Pittsburgh, PA (September 11–13 1990) pp 439–448
10 Olson, T J and Coombs, D J *Real-Time Vergence Control for Binocular Robots*, Technical Report 348, Department of Computer Science, University of Rochester (1990)
11 Hubel, D *Eye, Brain and Vision*, Scientific American Library, NY (1988)
12 Livingstone, M and Hubel, D 'Segregation of form, color, movement, and depth: Anatomy, physiology and perception', *Science*, Vol 240 (May 1988) pp 740–749
13 van Essen, D C, Felleman, D J, DeYoe, E A and Knierim, J J 'Probing the primate visual cortex: Pathways and perspectives', in A Valberg and B B Lee (eds.), *Advances in Understanding Visual Processes (Proc. NATO Advanced Research Workshop on "Advances in Understanding Visual Processes: Convergence of Neurophysiological and Psychophysical Evidence"*, Roros, Norway (August 6–10 1990)
14 Mallot, H A, von Seelen, W and Gianakopoulos, F 'Neural mapping and space-variant image processing', *Neural Networks*, Vol 3 (1990) pp 245–263
15 von Seelen, W and Mallot, H A 'Information processing in a neural architecture', in T Kohonen, K Mäkisara, O Simula and J Kangas (eds.), *Artificial Neural Networks (Proc. of 1991 Int. Conf. on Artificial Neural Networks (ICANN–91)*, Espoo, Finland (24–28 June 1991) Vol 1 p 855
16 Dow, B M 'Nested maps in macaque monkey visual cortex', in K N Leibovic (ed.), *The Science of Vision*, Springer, New York (1990) pp 84–124
17 Cavanagh, P 'Pathways in early vision', in Z W Pylyshyn (ed.), *Computational Processes in Human Vision: An Interdisciplinary Perspective*, Ablex, Norwood, NJ (1988) pp 239–261
18 Ingling Nr, C R and Grigsby, S S 'Perceptual correlates of magnocellular and parvocellular channels: Seeing form and depth in afterimages', *Vision Res.*, Vol 30 No 6 (1990) pp 823–828
19 LeVay, S, Hubel, D and Wiesel, T N 'The pattern of ocular dominance columns in macaque visual cortex revealed by a reduced silver stain', *J. Comp. Neur.*, Vol 159 (January 1975) pp 559–576
20 Aloimonos, J, Weiss, I and Bandopadhay, A 'Active vision', *Proc. 1st Int. Conf. on Comput. Vision*, London, UK (1987) pp 34–54
21 Aloimonos, J 'Purposive and qualitative active vision', *Proc. Image Understanding Workshop*, Pittsburgh, PA (September 11–13 1990) pp 816–828
22 Bandyopadhyay, A *A computational study of rigid motion perception*, PhD thesis, University of Rochester, Department of Computer Science (1986)
23 Tölg, S and Mallot, H A 'Tracking: ein Verfahren zur Stabilisierung bewegter Objekte mit einer aktiven Kamera', *Deutsche Arbeitsgemeinschaft für Mustererkennung (DAGM '90)*, Oberkochen-Aalen, Germany (1990) pp 642–649
24 Theimer, W and Mallot, H A 'Binocular vergence control and depth reconstruction using a phase method, *Deutsche Arbeitsgemeinschaft für Mustererkennung (DAGM '92)*, Dresden, Germany (1992) pp 133–140
25 Pahlavan, K and Eklundh, J-O 'A head-eye system – analysis and design' *Comput. Vision, Graph. & Image Process: Image Understanding* (July 1992) pp 41–56
26 Pahlavan, K, Uhlin, T and Eklundh, J-O 'Integrating primary ocular processes', *Second Euro. Conf. on Comput. Vision (ECCV-92)*, Santa Margherita Ligure, Italy (May 18–23 1992) pp 526–541
27 Burt P and Julesz, B 'A disparity gradient limit for binocular fusion', *Science*, Vol 208 (1980) pp 615–617
28 Jenkin, M R M and Jepson, A D 'The measurement of binocular disparity', in Z W Pylyshyn (ed.), *Computational Processes in Human Vision: An Interdisciplinary Perspectrive*, Ablex, Norwood, NJ (1988) pp 69–98
29 Jenkin, M R M, Jepson, A D and Tsotsos, J K 'Techniques for disparity measurement', *CVGIP: Image Understanding*, Vol 53 No 1 (1991) pp 14–30
30 Sanger, T D 'Stereo disparity computation using Gabor filters', *Biol. Cybern.*, Vol 59 (1988) pp 405–418
31 Yeshurun, Y and Schwartz, E L 'Cepstral filtering on a columnar imge architecture: A fast algorithm for binocular stereo segmentation', *IEEE Trans. PAMI*, Vol 11 No 7 (July 1989) pp 759–767
32 Bogert, B P, Healy, M J R and Tukey, J W 'The frequency analysis of time series for echoes: cepstrum, cross-cepstrum, and saphe cracking', *Proc. Symposium on Time Series Analysis* (1963) pp 209–243
33 Noll, A M 'Short-time spectrum and cepstrum techniques for vocal-pitch detection', *J. Acoust. Soc. Am.*, Vol 36 (1964) pp 296–302
34 Lee, D J, Krile, T F and Mitra, S 'Power cepstrum and spectrum techniques applied to image registration', *Appl. Opt.*, Vol 27 (1988) pp 1099–1106
35 Childers, D G, Skinner, D P and Kemerait, R C 'The cepstrum: A guide to processing', *Proc. IEEE*, Vol 65 (1977) pp 1428–1442
36 Bracewell, R N *The Hartley Transform*, Oxford University Press, Oxford (1986)
37 Kemerait, R C and Childers, D C 'Signal detection and extraction by cepstrum techniques', *IEEE Trans. Infor. Theory*, Vol 18 (91972) pp 745–759

38 Braitenberg, V 'Charting the visual cortex', in A Peters and E G Jones (eds.), *Cerebral Cortex*, Vol 3, Chap 10, Plenum Press, New York (1985) pp 379–414
39 Hubel, D H and Freeman, D C 'Projection into the visual field of ocular dominance columns in macaque monkey', *Brain Res.*, Vol 122 (1977) pp 336–343
40 Tootel, R B 'Deoxyglucose analysis of retinotopic organization in rimate striate cortex', *Science*, Vol 218 (1982) pp 902–904
41 Tyler, C W 'Spatial organization of binocular disparity sensitivity', *Vision Res.*, Vol 15 (1975) pp 583–590
42 Hassab, J C and Boucher, R 'A probabilistic analysis of time delay extraction by the cepstrum in stationary gaussian noise', *IEEE Trans. Infor. Theory*, Vol 22 (1976) pp 444–454
43 Lee, D J, Mitra, S and Krile, T F 'Dense depth map from 2-D cepstrum matching of image sequences', *Proc. Int. Workshop on Robust Comuter Vision*, Seattle, WA (October 1–3 1990) pp 200–207
44 Schwartz, E L *Novel Architectures for Imge Processing Based on Computer Simulation and Psychophysical Studies of Human Visual Cortex*, AFOSR – Technical Report 86–0059, New York University Medica Center (1986)
45 Ludwig, K-O, Neumann, B and Neumann, H 'Robust estimation of local stereoscopic depth', *Int. Workshop on Robust Computer Vision (IWRCV '92)*, Bonn, Germany (9–12 October 1992) pp 290–312
46 Ludwig, K-O, Neumann, B and Neumann, H 'Local stereoscopic depth estimation using ocular stripe maps', *Second Euro. Conf. on Comput. Vision (ECCV-92)*, Santa Margherita Ligure, Italy (1992) pp 373–377
46 Ludwig, K-O, Bergholm, F and Francisco, A 'The shape of the Horopter: a note', *Int. J. Comput. Vision*, Special issue on 'Stereopsis' (submitted 1993)

Counting people getting in and out of a bus by real-time image-sequence processing

F Bartolini*, V Cappellini* and A Mecocci†

The number of people getting in and out of a bus is an important parameter to allocate the proper number of buses for each connection-line of a public transport service. On the other hand, the correct distribution of the available buses over the different paths, is fundamental to obtain an optimization of the whole transport network, and to reduce costs. In this paper, an automatic system using dynamic image sequence processing to count people getting in and out of a bus is presented. Some fast algorithms are used to detect motion, estimate its direction, and validate the presence of moving people. The system can deal with vibrations, lighting fluctuations and environmental variations. The main advantages are the execution speed and the reliability of the counting process, is performed correctly even if people flow in a chaotic and very clustered way.

Keywords: dynamic sequence analysis, motion detection, target detection and counting, optimization of transport services

Efficient and cost-effective public transport services are very important. Some connection paths are usually provided by the public transport company; these paths link the most important areas of a city. A prearranged number of buses is allocated to each connection path; this number is different from path-to-path, and should vary during the day and from day-to-day in order to achieve optimum performance. The correct allocation of available buses to the different paths is a key point in obtaining efficient management of the bus fleet. This allocation cannot be done correctly without some efficient mechanism of gathering online information about the actual needs of each path. The number of passengers that are getting in and out of the buses can be used to estimate these needs. The increase of people-flow on a particular path in fact indicates that more buses should be available on that path. Of course, off-line statistical tests can be performed to get an idea of the average number of buses needed on the various paths. Nevertheless, the results that can be achieved with such tests are static in nature and can only be used for the general design of the whole transport network. The allocation can be optimized only with a 'real-time' reconfiguration strategy, due to the unavoidable temporal variation of the demand, and to its unpredictable random fluctuations.

*Dipartimento di Ingegneria Elettronica, Università di Firenze, via S. Marta 3, 50139 Firenze, Italy
†Dipartimento di Elettronica, Università di Pavia, via Abbiategrasso 209, 27100 Pavia, Italy

Paper received: 14 October 1992; revised paper received: 19 July 1993

SYSTEM OVERVIEW

In many applications, the problem of people counting has been attached by means of photoelectric sensors that give an *On/Off* response when a light beam is broken by an opaque target. These systems are not very reliable, especially when the flow is not regular and people pass close together. Photoelectric sensors cannot measure the actual cross-section of the moving target, and so spurious interruptions of the light beam, due to sudden events (e.g. a briefcase which can oscillate and periodically occlude the light beam) induce errors in the final count. Moreover, photoelectric sensors cannot easily determine the motion direction of the target, but this parameter is important to monitor the input *versus* the output flow.

Image sequence analysis allows us to discriminate among different false alarm conditions, and this gives more precise results. The proposed system alternates between two possible states: (a) target detection, and (b) target validation and direction-estimation. In the

0262–8856/94/010036–06