Multiresolution Estimation of 2-d Disparity Using a Frequency Domain Approach

A.D. Calway, H. Knutsson* and R. Wilson.

Dept. of Computer Science,
University of Warwick,
Coventry CV4 7AL, England.

* Computer Vision Laboratory,
Linköping University,
S-581 83 Linköping, Sweden.

Abstract

An efficient algorithm for the estimation of the 2-d disparity between a pair of stereo images is presented. Phase based methods are extended to the case of 2-d disparities and shown to correspond to computing local correlation fields. These are derived at multiple scales via the frequency domain and a coarse-to-fine 'focusing' strategy determines the final disparity estimate. Fast implementation is achieved by using a generalised form of wavelet transform, the multiresolution Fourier transform (MFT), which enables efficient calculation of the local correlations. Results from initial experiments on random noise stereo pairs containing both 1-d and 2-d disparities, illustrate the potential of the approach.

1 Introduction

Estimating the disparity between a pair of binocular images in order to determine depth information from a scene has received considerable attention for many years. Essentially a problem of finding corresponding points in the two views of the scene, the complexity of the task is considerable, involving not only the estimation of relative 2-d displacements, but with the added complication of taking into account such effects as geometric transformations and occlusions. This is reflected in the wide range of approaches to solving the problem that have been investigated [1, 2].

Recently, some of the problems have been successfully addressed by the use of frequency domain methods in the form of phase differencing. Using localised frequency representations similar to that proposed by Gabor [3], local phase differences between bandpass filtered versions of the binocular images provide robust estimation of disparity at sub-pixel accuracy [4, 5, 6]. Incorporation of the methods within some form of multiscale framework also allows for efficient matching to be achieved

BMVC 1992 doi:10.5244/C.6.24

using coarse-to-fine analysis [4, 6]. Nevertheless, these methods are not without their shortcomings. To the authors' knowledge, no straightforward extension to 2-d disparity estimation has been devised and the use of bandpass filters with constant relative bandwidth over scale would appear to be an unwelcome restriction: significant events in a scene are in general broadband and in any case there is no reason why the disparity between the views of a given object should be directly linked to its size or frequency content. In addition, the technique is critically dependent upon the local frequency properties of the images - dominant frequencies significantly different from that of the filter centre frequency can lead to error.

The work reported here is an attempt to address the problems of phase differencing while retaining the advantages of a frequency domain approach. The estimation of 2-d disparity is cast in the form of a least squares minimisation problem over all spatial frequencies and is shown to be equivalent to calculating spatial correlations via the frequency domain. Moreover, by defining the scheme within the framework of a generalised wavelet transform, the multiresolution Fourier transform (MFT) [7, 8], the analysis can be based upon local spatial regions over a range of sizes and so enable a fast coarse-to-fine matching strategy to be adopted. This approach avoids the problem of tying disparity to scale and removes the dependence on a known centre frequency associated with phase differencing. In addition, these advantages are achieved without the high computational cost normally associated with correlation methods. Finally, the unified framework provided by the MFT gives potential for extending the scheme to incorporate feature information to further aid in guiding the disparity estimation and to cope with problems such as local geometric transformations. After outlining the theoretical principles of the algorithm and its implementation using the MFT, results of experiments on random noise stereo pairs with 1-d and 2-d disparities are presented to illustrate the potential of the approach.

2 Multiresolution Disparity Estimation

The purpose of this section is to outline the main features of the algorithm and to indicate its relationship with phase based methods. Towards this end, consider a 2-d image $x(\vec{\xi})$ at a depth Δ from the reference (vergence) plane in a binocular system, where $\vec{\xi} = (\xi_1, \xi_2)$ is the coordinate vector in 2-d space. Ignoring any effects such as scaling, the left and right images in the system are related by

$$x_R(\vec{\xi}) = x_L(\vec{\xi} + \vec{d}) \tag{1}$$

where the 2-d disparity $\vec{d}$ is proportional to the depth Δ. This relationship can also be considered in the Fourier domain as

$$\hat{x}_R(\vec{\omega}) = \hat{x}_L(\vec{\omega}) \exp[\jmath \vec{\omega}.\vec{d}] \tag{2}$$

where '.' denotes scalar product, $\jmath = \sqrt{-1}$ and $\hat{x}(\vec{\omega})$ is the 2-d Fourier transform (FT) of $x(\vec{\xi})$. The significance of (2) is that it suggests a means of estimating $\vec{d}$ using the phase of the inner product between $\hat{x}_L(\vec{\omega})$ and $\hat{x}_R(\vec{\omega})$, ie

$$\arg[\hat{x}_L(\vec{\omega})\hat{x}_R^*(\vec{\omega})] = -(\vec{\omega}.\vec{d}) \tag{3}$$

Thus, providing the direction of $\vec{d}$ is known (eg when considering horizontal disparities) and spectral estimates of the left and right images are obtained at some known frequency $\vec{\omega}$, then estimation of $\vec{d}$ can be made using (3). This is the basis of phase differencing approaches. However, the situation is less clear when the direction of disparity is unknown, as is the case in most natural stereopsis problems [9, 10]. In this instance, single frequency estimates will mean that $\vec{d}$ is indeterminate; to determine $\vec{d}$ requires more than one frequency estimate in different radial directions. This then poses the question as to what is the most appropriate way of estimating $\vec{d}$: to use a subset of frequencies or to devise a method using all frequencies. Given that in most natural scenes it is impossible to predict a priori in which frequency bands significant events will lie, it would seem that the latter should be the preferred option. In fact, such a method of solution can be readily formulated in terms of a least-squares problem and corresponds to selecting $\vec{d}$ to maximise the function

$$\rho(\vec{d}) = \mathcal{F}^{-1}[\hat{x}_L(\vec{\omega})\hat{x}_R^*(\vec{\omega})] = \frac{1}{4\pi^2}\int_{-\infty}^{\infty} \hat{x}_L(\vec{\omega})\hat{x}_R^*(\vec{\omega}) \exp[j\vec{\omega}.\vec{d}]\, d\vec{\omega} \tag{4}$$

where $\mathcal{F}^{-1}$ denotes the inverse 2-d FT. The maximisation therefore amounts to finding the 'phase correction' term $(\vec{\omega}.\vec{d})$ which maximises the inner product between the spectra of the binocular images, ie $\arg[\hat{x}_R^*(\vec{\omega})]$ is 'rotated' so as to minimise the squared error between the spectra. Moreover, (4) can be written in terms of the spatial domain as [11]

$$\rho(\vec{d}) = \int_{-\infty}^{\infty} x_L(\vec{\xi})x_R(\vec{\xi}+\vec{d})\, d\vec{\xi} \tag{5}$$

which is just the cross correlation between $x_L(\vec{\xi})$ and $x_R(\vec{\xi})$. Maximising (4) therefore corresponds to finding the peak in the correlation field and the connection between phased based methods and correlation is made clear: the latter provides a natural extension of the former to deal with 2-d disparities, by making use of the whole frequency domain.

Of course, simply computing the global correlation between the left and right images is inappropriate except in the most trivial of cases. In practice, the interocular disparity will be inherently local; objects in a scene are necessarily confined to some finite spatial region and exist at differing depths, implying that any correspondence measurements must also be based on local properties [6]. This can be achieved in the present case by considering local correlations between neighbourhoods in the binocular images, ie find the $\vec{d}$, denoted by $\vec{d}(\vec{\xi}_1, \vec{\xi}_2)$, which maximises

$$\rho(\vec{\xi}_1, \vec{\xi}_2, \vec{d}) = \frac{1}{4\pi^2}\int_{-\infty}^{\infty} \hat{x}_L(\vec{\xi}_1, \vec{\omega})\hat{x}_R^*(\vec{\xi}_2, \vec{\omega}) \exp[j\vec{\omega}.\vec{d}]\, d\vec{\omega} \tag{6}$$

where the global spectra in (4) are now replaced by the local spectra $\hat{x}_L(\vec{\xi}_1, \vec{\omega})$ and $\hat{x}_R(\vec{\xi}_2, \vec{\omega})$, centred at $\vec{\xi}_1$ and $\vec{\xi}_2$ respectively, and defined according to

$$\hat{x}(\vec{\xi}, \vec{\omega}) = \int_{-\infty}^{\infty} w(\vec{\chi} - \vec{\xi})x(\vec{\chi}) \exp[-j\vec{\omega}.\vec{\chi}]\, d\vec{\chi} \tag{7}$$

where $w(\vec{\xi})$ is some appropriate window function, ie $\hat{x}(\vec{\xi}, \vec{\omega})$ is a windowed FT reminiscent of the Gabor representation [3]. It is these equations which underlie the disparity estimation used in the present work: derive local correlation fields $\rho(\vec{\xi}_1, \vec{\xi}_2, \vec{d})$

by computing the inverse FT of $\hat{x}_L(\vec{\xi}_1, \vec{\omega})\hat{x}_R^*(\vec{\xi}_2, \vec{\omega})$ and find the peak to give the disparity $\vec{d}(\vec{\xi}_1, \vec{\xi}_2)$.

The above formulation also suggests a means of overcoming the matching problem, ie selecting $\vec{\xi}_1$ and $\vec{\xi}_2$ in (6). If the neighbourhoods used in the correlations are too small, then finding the best match will involve extensive searching, whereas neighbourhoods which are too large will be susceptible to error due to the presence of more than one disparity. As has been previously noted, eg in [4, 6], the solution is to employ some form of coarse-to-fine analysis so that the disparity estimates can be 'focused' over multiple scales. This approach can be incorporated here by defining the local correlations to be dependent upon a scale parameter σ, ie

$$\rho(\vec{\xi}_1, \vec{\xi}_2, \sigma, \vec{d}) = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \hat{x}_L(\vec{\xi}_1, \vec{\omega}, \sigma)\hat{x}_R^*(\vec{\xi}_2, \vec{\omega}, \sigma) \exp[j\vec{\omega}.\vec{d}]\, d\vec{\omega} \tag{8}$$

where the local spectra are now scale dependent and correspond to multiresolution Fourier transforms (MFT) [8, 7]

$$\hat{x}(\vec{\xi}, \vec{\omega}, \sigma) = \sigma \int_{-\infty}^{\infty} w(\sigma(\vec{\chi} - \vec{\xi}))x(\vec{\chi}) \exp[-j\vec{\omega}.\vec{\chi}]\, d\vec{\chi} \tag{9}$$

ie a 'stack' of windowed FTs in which the locality of the spectral estimates is varied as a function of σ. Using (8), it is therefore possible to derive local correlation fields, and thus disparity estimates, at multiple spatial resolutions via the Fourier domain. Moreover, there is now no longer a link between those disparity estimates and a specific frequency band as in previous multiscale approaches; in this case, the estimates are based on information from the whole of the frequency domain.

It is now possible to summarise the multiresolution scheme used to derive the required disparity field. Starting at some suitably large scale σ_0, local correlations between neighbourhoods centred at the same spatial positions in the left and right images are formed according to (8) and the disparities $\vec{d}(\vec{\xi}, \vec{\xi}, \sigma_0)$ found which maximise $\rho(\vec{\xi}, \vec{\xi}, \sigma_0, \vec{d})$. A disparity field $\vec{D}(\vec{\xi}, \sigma_0)$ is then generated such that $\vec{D}(\vec{\xi}, \sigma_0) = \vec{d}(\vec{\xi}, \vec{\xi}, \sigma_0)$, ie it represents the current disparity estimate with respect to the left image at spatial position $\vec{\xi}$ and scale σ_0. The scheme then proceeds through smaller and smaller scales ($\sigma_0 < \sigma_1 ... < \sigma_{m-1} < \sigma_m$), deriving disparity fields at each scale according to the following update rule

$$\vec{D}(\vec{\xi}, \sigma_{k+1}) = \vec{D}(\vec{\xi}, \sigma_k) + \vec{d}(\vec{\xi}, \vec{\xi} + \vec{D}(\vec{\xi}, \sigma_k), \sigma_{k+1}) \qquad 0 \leq k < m \tag{10}$$

where the first term on the rhs serves as both the previous estimate and the 'focusing' term - defining the pair of regions to be correlated - and the second term is the disparity update at scale σ_{k+1} based on the current correlation. The local correlations performed at smaller scales are therefore directed by the disparity estimates obtained at larger scales, producing a more refined estimate at each stage. The final estimate is then given by the disparity field $\vec{D}(\vec{\xi}, \sigma_m)$ defined at scale σ_m.

3 Implementation

3.1 The Discrete MFT

The algorithm described above is based upon the MFT as defined by (9). This is a generalised form of wavelet transform designed specifically to enable local Fourier

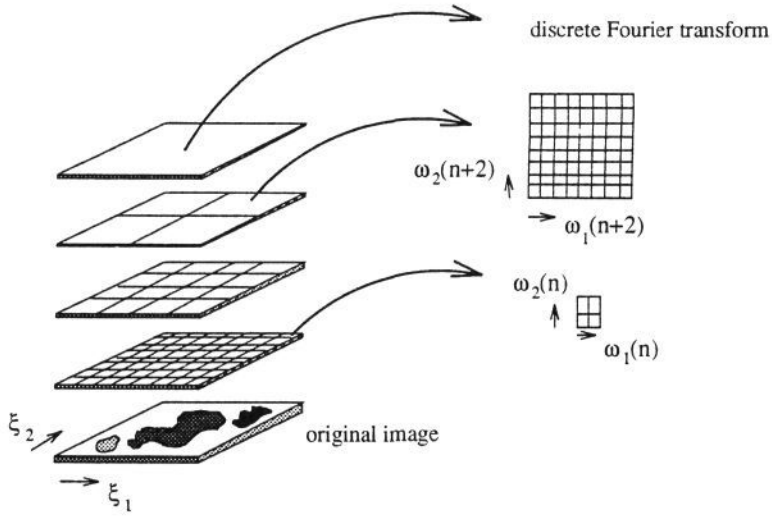


Figure 1: The MFT viewed as a quadtree in which the nodes are assigned the local spectra corresponding to the region "below" the node.

analysis to be performed at multiple scales [8]. A brief summary of the essential properties of the discrete transform is given here. For a discrete 2-d image $x(\vec{\xi}_i)$, its MFT at scale $\sigma(n)$, frequency $\vec{\omega}_j(n)$ and position $\vec{\xi}_i(n)$ is given by

$$\hat{x}(\vec{\xi}_i(n), \vec{\omega}_j(n), \sigma(n)) = \sum_k w_n(\vec{\xi}_k - \vec{\xi}_i(n)) x(\vec{\xi}_k) \exp[-j\vec{\xi}_k.\vec{\omega}_j(n)] \qquad (11)$$

where the discrete window sequence $w_n(\vec{\xi}_i)$ approximates a scaled version of a suitable continuous function $w(\vec{\xi})$, ie $w_n(\vec{\xi}_i) = \sigma(n)w(\sigma(n)\vec{\xi}_i)$. Thus, for some value of $\sigma(n)$, $\hat{x}(\vec{\xi}_i(n), \vec{\omega}_j(n), \sigma(n))$ is a discrete windowed FT of $x(\vec{\xi}_i)$ and corresponds to local frequency estimates centred at spatial positions $\vec{\xi}_i(n)$. As $\sigma(n)$ varies, the spatial and frequency resolution varies, and thus the transform as a whole consists of local estimates over a range of scales.

The two most important factors determining the properties of the MFT are the distribution of the sampling points $\vec{\xi}_i(n)$ and $\vec{\omega}_j(n)$, and the choice of the window sequence. In the present work, the 2-d transform has been formed as the cartesian product of 1-d transforms, and the sampling points in both domains distributed on regularly spaced square lattices of size $N_\xi(n) \times N_\xi(n)$ and $N_\omega(n) \times N_\omega(n)$, where $N_\xi(n)N_\omega(n) = 2N$ for an image of finite size $N \times N$ [8]. The window functions adopted here are bandlimited versions of the prolate spheroidal sequences [11]. These provide maximal spatial localisation and enable efficient computation of the transform using fast Fourier transform techniques [7]. A useful interpretation of the resulting transform is that of a quadtree structure in which the individual nodes are assigned the local spectra referring to the neighbourhood "below" the node and have four associated child nodes whose estimates refer to quadrants of the father's neighbourhood (see Fig. 1). It is this hierarchical framework which forms the basis of the disparity focusing algorithm described below.

3.2 Disparity Focusing

The basic operation employed in the disparity estimation can now be expressed in terms of the discrete MFT coefficients of two binocular images, ie (cf (8))

$$\rho(\vec{\xi}_i(n), \vec{\xi}_k(n), \sigma(n), \vec{d}) = F^{-1}_{N_\omega(n)} \left[\hat{x}_L(\vec{\xi}_i(n), \vec{\omega}_j(n), \sigma(n)) \hat{x}^*_R(\vec{\xi}_k(n), \vec{\omega}_j(n), \sigma(n)) \right] \quad (12)$$

where $F^{-1}_{N_\omega(n)}$ denotes the inverse 2-d discrete FT of size $N_\omega(n) \times N_\omega(n)$ and the correlation is performed between neighbourhoods centred at $\vec{\xi}_i(n)$ and $\vec{\xi}_k(n)$ in the left and right images respectively. A correlation field of size $N_\omega(n) \times N_\omega(n)$ is thereby obtained, in which the position of the peak indicates the relative 2-d displacement between the two neighbourhoods. As a guide to the computational saving obtained by using the MFT, the computational burden of implementing (12) for an $N \times N$ pixel image is in the order of $4N^2_\omega(n) \log_2 2N$ multiplications [7], giving a gain by a factor of around $N^2_\omega(n)/4 \log_2 2N$ over that required for direct calculation of the correlation. For example, for a 256×256 image and a 16×16 neighbourhood, this corresponds to a gain by a factor greater than 5, whilst for neighbourhoods of 32×32 and 64×64, this increases to over 25 and 100 respectively. The saving achieved is therefore considerable, particularly at the larger region sizes.

The disparity focusing algorithm is best described in terms of the quadtree framework discussed above. A level of the MFT, $n = n_0$ say, is chosen as the starting level (typically corresponding to $N_\omega(n) = 64$) and either the left or right channel selected as the reference channel. The algorithm then proceeds as follows (Fig. 2):

1. Cross correlations between corresponding nodes on level n_0 are formed and peak positions in the correlation fields assigned to the relevant nodes in the reference channel.

2. For a father node in the reference channel on level n_0, its child nodes at level n_0+1 are compared with those on the same level in the other channel according to the disparity estimate at the father node. If the estimate is greater than half a block at level $n_0 + 1$ along either or both coordinates, the child nodes are compared with their relevant "neighbours" in the other channel; otherwise they are compared with their corresponding nodes. The peak positions in the resulting correlation fields are used to produce an updated estimate (cf (10)), which is then assigned to the relevant nodes on level $n_0 + 1$ of the reference channel.

3. The process proceeds to level $n_0 + 2$, nodes are compared according to the disparities obtained at the previous level (ie to the nearest block interval) and a new disparity estimate produced. This process then continues through subsequent levels until some final level $n_0 + m$ is reached.

The result of this hierarchical scheme is a set of disparity estimates defined at levels $n_0 \leq n \leq n_0 + m$, with the spatial resolution of each estimate being determined by the corresponding resolution of the MFT level.

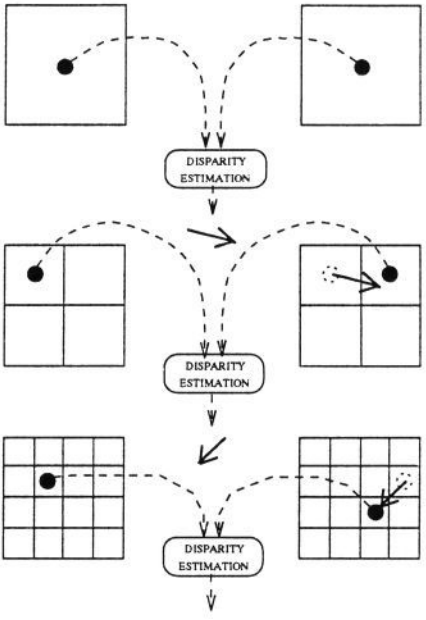


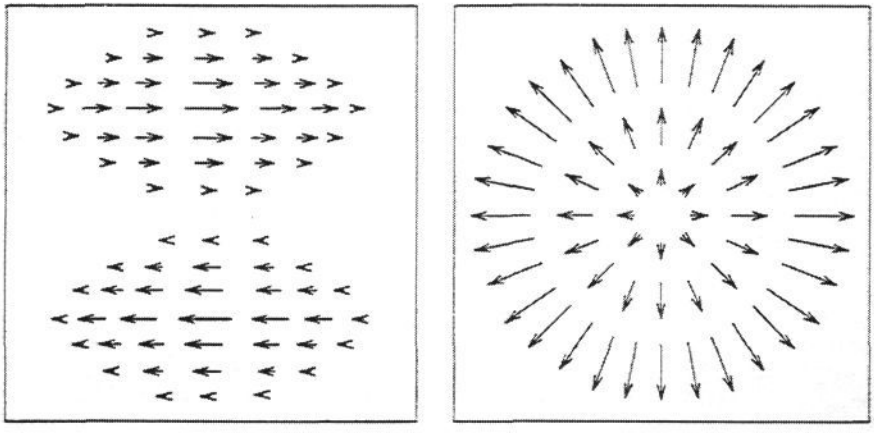
Figure 2: Discrete disparity focusing within the quadtree framework. Estimates obtained at higher nodes dictate which nodes are compared at subsequent levels, yielding an updated and refined disparity field at each scale.

4 Experiments

To test the algorithm, experiments were performed on random noise stereo pairs with horizontal and 2-d disparities. The images were of size 256×256 pixels with 8-bit grey level resolution. The MFTs of each image were generated and the levels with $16 \leq N_\omega(n) \leq 64$ used in the focusing algorithm.

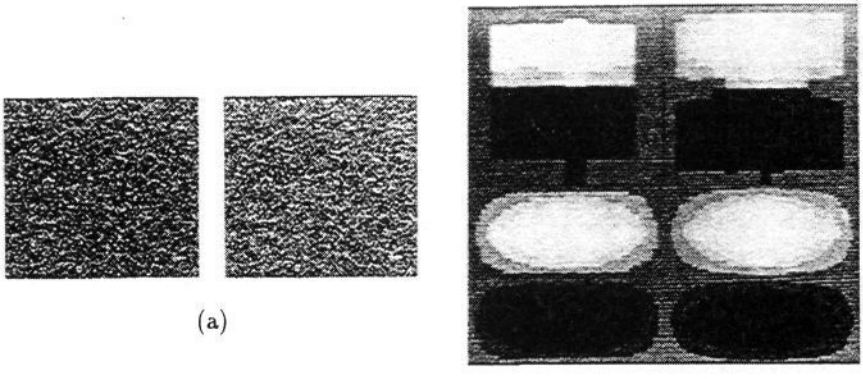
The test image pairs are shown in Figs. 4a and 5. The first of these consists of only horizontal disparities, linearly increasing in the positive and negative directions to a peak of ± 16 pixels in the centre of the upper and lower halves of the image respectively, ie forming inward and outward projecting peaks when viewed stereoscopically (Fig. 3a). The second pair incorporates 2-d disparities by varying the relative displacement as a function of the radius from the centre of the image. where the disparity at the edges is 16 pixels (Fig. 3b).

Results of the experiments are shown in Figs. 4b and Fig. 6. These show the horizontal and vertical components of the estimates obtained on each of the four levels (only the horizontal component in the case of the first pair, the vertical component being zero). The luminance values (0-255 grey levels) in these images indicates the amount of disparity, where zero disparity corresponds to a grey level value of 128. These results show clearly the focusing steps of the algorithm and the final estimates correspond well to the known disparity variation.



(a) (b)

Figure 3: Test image disparity variation with (a) horizontal disparities and (b) 2-d disparities.



(a)

(b)

Figure 4: (a) Random noise stereo pair containing horizontal disparities. (b) Horizontal component of disparity estimates produced by focusing algorithm from the stereo pair in (a).

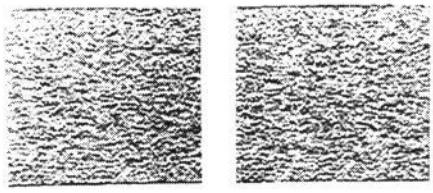
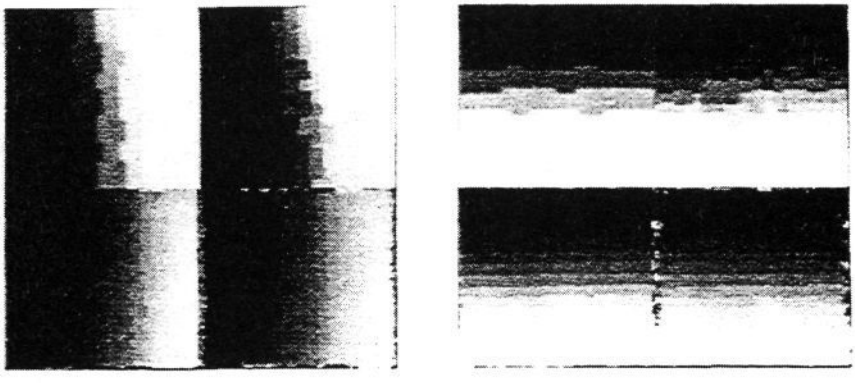


Figure 5: Random noise stereo pair containing 2-d disparities.



(a) (b)

Figure 6: (a) Horizontal and (b) vertical component of disparity estimates produced by focusing algorithm from the stereo pair in Fig. 5.

5 Conclusions

An algorithm to compute the 2-d disparity between a pair of binocular images has been presented. The approach is based on the calculation of local correlation fields over multiple scales using a frequency domain method. This has been shown to be a natural extension of phase differencing techniques to deal with 2-d disparities. Efficient implementation of the algorithm is achieved by making use of the MFT. A disparity focusing scheme enables fast matching of corresponding regions in the two images and the results obtained from experiments illustrate the satisfactory performance of the approach.

It should be emphasized, however, that the work presented here is in its preliminary stages. The initial experiments suggest that the approach has considerable potential, although the simplicity of particularly the matching will clearly lead to difficulties when dealing with more complex scenes. Work is under way in extending the approach, most notably on incorporating both local transformation and feature information into the algorithm. Perhaps the most interesting aspect of this work is that due to the flexibility and richness of representation provided by the MFT, the potential exists for incorporating such extensions within the same framework [8].

References

[1] S.T.Barnard and M.A.Fischler, Computational stereo, ACM Computing Surveys 1982; 14, 4: 553-572.

[2] U.R.Dhond and J.K.Aggarwal, Structure from stereo - a review, IEEE Trans. Sys., Man and Cybern. 1989; 19, 6: 1489-1510.

[3] D.Gabor, Theory of communication, Proc. IEE 1946; 93: 429-441.

[4] A.D.Jepson and D.J.Fleet, Fast computation of disparity from phase differences, In: Proc. IEEE Conf. Comput. Vision Patt. Recog., San Diego, 1989, pp 398-403.

[5] T.Sanger, Stereo disparity computation using Gabor filters, Biol. Cybern 1988; 59: 405-418.

[6] R.Wilson and H.Knutsson, A multiresolution stereopsis algorithm based on the Gabor representation, In: Proc. IEE Int. Conf. Image Process. & its Appl., Warwick, 1989, pp 19-22.

[7] A.D. Calway, The Multiresolution Fourier Transform: A General Purpose Tool for Image Analysis, Ph.D. Thesis, Warwick University, 1989.

[8] R.Wilson, A.D.Calway, and E.R.S.Pearson, A generalised wavelet transform for Fourier analysis: the multiresolution Fourier transform and its application to image and audio signal analysis, IEEE Trans. Inform. Th. 1992; 38, 2: 674-690.

[9] H.C.Longuet-Higgins, The role of vertical dimension in stereoscopic vision, Perception 1982; 11: 377-386.

[10] J.Mayhew, The interpretation of stereo-disparity information: the computation of surface orientation and depth, Perception 1982; 11: 387-403.

[11] A. Papoulis, Signal Analysis, McGraw-Hill, New York, 1977.