Multiresolution stereo image matching using complex wavelets

Julian Magarey and Anthony Dick
Cooperative Research Centre for Sensor Signal and Information Processing
SPRI Building, Technology Park Adelaide, SA 5095, Australia
Email: jfam@cssip.edu.au, anthony@cs.adelaide.edu.au

Abstract

This paper describes a multiresolution image-matching strategy, based on the Complex Discrete Wavelet Transform (CDWT), to derive a dense disparity field with hierarchical (coarse-to-fine) refinement. The CDWT feature space efficiently provides fractionally accurate matching results which are robust to typical image formation perturbations such as offsets, global scaling, and additive noise. At each level of the hierarchy, the disparity field is regularised *to provide a global compromise between feature similarity and disparity field continuity, resulting in feature-sensitive smoothing. The algorithm is well suited to analysing facial images, for which we demonstrate striking reconstruction results.*

1. Introduction

This paper describes a depth reconstruction algorithm for binocular stereo which uses a dense multiresolution matching strategy. Multiresolution strategies have been shown to be useful in resolving the ambiguities of dense matching smooth surfaces which lack clear feature points and contain few if any occlusions [5]. The matching is fully two-dimensional as we do not make use of the epipolar constraint to reduce the dimensionality of the problem. The key feature of our algorithm is the feature space in which matching takes place, which is based on the Complex Discrete Wavelet Transform [6]. It has been demonstrated that this feature space yields matching results which are more robust than original-pixel-intensity-domain matching to typical distortions of binocular stereo pairs such as level shifts, scaling, and additive noise. Even so, various sources of error can corrupt the disparity field, particularly over broad expanses with little texture or detail. For this reason, following Anandan [1], we insert a *regularisation* step after matching at each level, which trades off feature space similarity for disparity field smoothness. Our matching procedure provides a directional *confidence measure* which assists this regularisation step, preserving sharp features while smoothing over untextured regions.

We chose to demonstrate the performance of our algorithm on faces for several reasons. Firstly, face recognition is a significant application area in its own right, and 3-d reconstructions can be a key part of this. Secondly, as faces consist mostly of smooth untextured surfaces with relatively few strong features, they present severe problems for feature-based matching algorithms, but are well suited to our dense matching algorithm. Thirdly, the human visual system is strongly adapted to analysing facial structure, drawing the eye to any areas which do not “look right”. Thus we have a ready-made mechanism for qualitatively assessing the performance of our algorithm in the absence of ground truth values.

2. Outline of our algorithm

Most previous matching strategies have been based on cross-correlation in the intensity domain [3]. While this achieves satisfactory results in certain cases, it is usually implemented using a computationally intensive exhaustive search. It has been shown that *phase-based* methods of disparity measurement [4] can overcome some of the weaknesses of intensity-domain correlation in stereo matching. The principle is based on the the Fourier shift theorem, which relates global translations in the spatial domain to phase rotations in the frequency domain. Application of this theorem to the coefficients of a *local* Fourier basis expansion can yield a spatially varying field of disparity estimates which is robust to certain image formation perturbations including level shifts and intensity scaling. Gabor basis functions (Gaussian-windowed complex exponentials) are often used because of their optimum combined compactness in time and frequency domains [2].

2.1. The Complex Discrete Wavelet Transform

Our aim was to implement a multiresolution phase-based matching algorithm. The separable discrete wavelet transform (DWT) [7] is an efficient recursive structure for producing a multiresolution analysis of digital images. The DWT is implemented using repeated applications of a simple processing block consisting of 1-d convolutions with an FIR basis pair, and downsampling operations. To obtain the phase information, we used the Complex Discrete Wavelet Transform (CDWT) [6], a variant of the DWT based on a complex-valued FIR filter pair with Gabor-like characteristics. The transform is efficiently implementable using only integer arithmetic, as the FIR filters are rational-valued with length 4.

The output at level m consists of six complex subimages $\{D^{(n,m)}, n = 1, \ldots, 6\}$. It may be shown [6] that each subimage may be regarded as the (downsampled) output of an equivalent *wavelet filter* which has 2-d Gabor form. Each wavelet filter, and its corresponding subimage, has a characteristic orientation specified by the spatial frequency $\mathbf{\Omega}_{n,m}$. The basis pair was designed so that the six orientations are evenly spaced between $-\pi/2$ and $\pi/2$, among other desirable properties [6].

Because of the downsampling (by 2^m at level m), the subimages consist of complex coefficients defined on a *subpixel* grid, in which each subpixel subtends a region of pixels in the original image. The size of the region, and therefore the scale of the features being extracted, increases with m.

2.2. Matching in the complex wavelet domain

The CDWT is applied to both left and right images. We define our CDWT-domain matching criterion between level m left and right subpixels $\mathbf{x}$ and $\mathbf{x}'$ as follows:

$$SD^{(m)}(\mathbf{x}, \mathbf{x}') = \sum_{n=1}^{6} \left| D_l^{(n,m)}(\mathbf{x}) - D_r^{(n,m)}(\mathbf{x}') \right|^2 \quad (1)$$

The correspondent for the left subpixel $\mathbf{x}$ is taken to be the right subpixel $\mathbf{x}'$ which minimises this quantity. The level m left-right disparity $\mathbf{d}_m(\mathbf{x})$ is $\mathbf{x}' - \mathbf{x}$. (Because of the downsampling, the disparity over the region of pixels subtended by $\mathbf{x}$ is $2^m \mathbf{d}_m(\mathbf{x})$.)

In a coarse-to-fine strategy, it is important that the matching subpixel is located to fractional accuracy. The Gabor-like nature of the wavelet filters allows us to accurately estimate coefficient values in between the subpixel grid, by interpolation from the surrounding grid coefficients [6]. Hence, given a starting subpixel $\mathbf{x}'$ for each $\mathbf{x}$, we can use the coefficient $D_l^{(n,m)}(\mathbf{x})$ and those surrounding $D_r^{(n,m)}(\mathbf{x}')$ to estimate a *matching surface* consisting of the values of $SD^{(m)}(\mathbf{x}, \mathbf{x}' + \mathbf{f})$ over a real 2-d interval of offsets $\mathbf{f}$. Over small offsets, a 2-d quadratic model fits the surface well:

$$SD^{(m)}(\mathbf{x}, \mathbf{x}' + \mathbf{f}) \approx \frac{1}{2}(\mathbf{f} - \mathbf{f}_0)^T \mathcal{K} (\mathbf{f} - \mathbf{f}_0) + \delta \quad (2)$$

The surface parameters $\{\mathbf{f}_0, \mathcal{K}, \delta\}$ (the surface minimum location, curvature matrix, and minimum value) may be computed directly from the coefficients $D_l^{(n,m)}(\mathbf{x})$ and $D_r^{(n,m)}(\mathbf{x}')$ and the spatial frequencies $\mathbf{\Omega}_{n,m}$. It turns out that $\mathbf{f}_0$, and hence the level m disparity, depends heavily on the *phase* of the two complex coefficient vectors, while the magnitudes have comparatively little influence. The dependence of disparity on phase and insensitivity to magnitude gives the algorithm immunity to global perturbations such as level shifts and intensity scaling. This means that the matching is based on a generalised level-and-scaling-independent area similarity, increasing the robustness of the algorithm.

The shape of the matching surface $SD^{(m)}$, encapsulated by the symmetric 2×2 matrix $\mathcal{K}$, acts as our directional measure of confidence. It may be shown [6] that the curvature of the surface in a given direction indicates our relative confidence in the component of $\mathbf{f}_0$ in that direction.

2.3. Regularisation of the disparity field

Even with its robustness, the disparity field is subject to random errors. If these were allowed to propagate down the pyramid, the final reconstructed surface would be extremely jagged. To minimise this risk, we use a regularisation procedure based on that proposed by Anandan in [1]. Given the field $\{\mathbf{f}_0\}$ of feature space minimum location offsets, we aim to obtain an optimal compromise between feature similarity and disparity field smoothness by finding a field $\{\mathbf{u}\}$ which minimises the functional

$$E(\{\mathbf{u}\}) = E_{sm}(\{\mathbf{u}\}) + \lambda E_{ap}(\{\mathbf{u}\}) \quad (3)$$

The term $E_{ap}(\{\mathbf{u}\})$ is the global "approximation error" energy, which we define as the normalised sum of the differences between $SD(\mathbf{x}, \mathbf{x}' + \mathbf{u})$ and the feature-space minimum value δ:

$$E_{ap}(\{\mathbf{u}\}) = \sum_{\mathbf{x}} \frac{SD^{(m)}(\mathbf{x}, \mathbf{x}' + \mathbf{u}) - \delta}{\delta} \quad (4)$$

The smoothness term $E_{sm}(\{\mathbf{u}\})$ is defined in terms of the gradients of the components (u_x, u_y) of $\mathbf{u}$:

$$E_{sm}(\{\mathbf{u}\}) = \sum_{\mathbf{x}} \|\nabla u_x\|^2 + \|\nabla u_y\|^2 \quad (5)$$

The scaling factor λ in (3) controls the relative influence of each term on the resultant disparity field. The smaller it is made, the greater the smoothing effect of the regularisation.

An exact global minimisation of $E(\{\mathbf{u}\})$ is computationally expensive to compute. A good approximate solution my be found using only the local information surrounding each subpixel, by solving

$$(\mathbf{u} - \overline{\mathbf{u}}) + \lambda c_{max} \Delta \mathbf{u}_{max} + \lambda c_{min} \Delta \mathbf{u}_{min} = \mathbf{0} \quad (6)$$

where $\Delta\mathbf{u}_{max}$ and $\Delta\mathbf{u}_{min}$ are the components of $\mathbf{u} - \mathbf{f}_0$ along the (orthogonal) principal directions of $SD^{(m)}$, c_{max} and c_{min} are its principal curvatures (normalised by δ), and $\overline{\mathbf{u}}$ is the four-neighbour average of $\mathbf{u}$. Equation (6) can be solved using Gauss-Seidel iteration; see [1] for details. After convergence over the whole image, the smoothed disparity estimates $\overline{\mathbf{d}}_m(\mathbf{x})$ are set to $\mathbf{x}' - \mathbf{x} + \mathbf{u}$.

The presence of the matching surface curvatures in the regularisation acts to reduce its smoothing effect in regions and directions in which the confidence of the match is high, while increasing its effect in regions and directions of lower confidence. This results in the preservation of sharp features and the effective smoothing of disparity over featureless surfaces, where estimation is dominated by noise. Further improvement was obtained by thresholding, before regularisation, the field $\{\mathbf{f}_0\}$ based on the product $c_{max}c_{min}$, which is an effective scalar measure of confidence.

2.4. Coarse-to-fine refinement

Matching begins at the coarsest level m_{max} with a starting guess of zero, effected by setting $\mathbf{x}' = \mathbf{x}$. The smoothed disparity field $\{\overline{\mathbf{d}}_{m_{max}}\}$ is used as starting guesses for estimation at the next finer level $m_{max} - 1$, after scaling up by 2 to allow for the denser spacing of subpixels. We use the *overlapping pyramid* scheme of Anandan [1], in which the disparities at the four level $m+1$ subpixels directly subtending the level m subpixel under scrutiny are all used as candidate starting guesses. The candidate which leads to the best match (lowest δ value) is accepted for the regularisation step. Estimation proceeds in this fashion, decrementing m from coarse to fine, with matching followed by regularisation, until the finest level of detail ($m = 1$) is reached.

3. Results

The tests were carried out using a pair of PULNiX-CN5 cameras, with a horizontal baseline of approximately 3 cm. The images were digitised as 8-bit greyscale, and cropped to dimensions 572 by 768. Seven levels of CDWT were applied to each, so that the coarsest level disparity field consisted of 4 by 6 estimates. We found that setting λ to 0.25 gave an appropriate combination of smoothing and feature preservation for this data set, and that no more than 10 Gauss-Seidel iterations were required at each level to obtain convergence over the whole image. The final disparity field (of size 286 by 384) was combined with the previously ascertained camera calibration matrices in a triangulation procedure [3] to produce a dense depth map. The total processing time was approximately ten minutes on a Sun Ultra workstation.

Results obtained from two different face image pairs are shown in Figure 1. They are given a realistic appearance by mapping the left image onto the surface formed by fitting planes to adjacent triplets of points. The reconstructions are smooth in the cheek and forehead regions while maintaining good clarity of features, indicating that regularisation has been feature-sensitive as designed. Furthermore the scalar confidence threshold was successful in removing the influence of noise-dominated background estimates on the regularisation step.

4. Conclusion

In this paper we have described a matching algorithm which is suitable for producing dense depth maps for smooth and relatively featureless surfaces from stereo image pairs. The algorithm is multiresolution in structure and is based on similarity in the Complex Discrete Wavelet Transform domain. The properties of this feature space allow efficient fractionally-accurate matching at each level, in a manner robust to global intensity perturbations such as scaling, offsets, and additive noise. Disparity estimates are accompanied by an intuitively meaningful directional confidence measure. The confidence measure is utilised in a regularisation algorithm which is carried out at each level of processing to smooth away noise-induced disparity errors in relatively featureless regions while preserving disparities based on clear features. The effectiveness of the algorithm has been demonstrated in 3-d reconstruction of faces from calibrated stereo image pairs.

References

[1] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *Intern. J. Comput. Vis.*, 2:283–310, 1989.

[2] J. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A*, 2(7):1160–1169, July 1985.

[3] O. Faugeras. *Three-dimensional computer vision*. MIT Press, 1993.

[4] D. Fleet, A. Jepson, and M. Jenkin. Phase-based disparity measurement. *CVGIP: Image Understanding*, 53(2):198–210, March 1991.

[5] T. Lindeberg. Direct estimation of affine image deformations using visual front-end operations with automatic scale selection. In *Proc. 5th International Conference on Computer Vision*, pages 134–141, June 1995.

[6] J. Magarey and N. Kingsbury. Motion estimation using a complex-valued wavelet transform. *IEEE Transactions on Signal Processing Special Issue on Theory and Applications of Filter Banks and Wavelet Transforms*, April 1998. *(To appear)*.

[7] S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. on PAMI*, 11(7):674–693, July 1989.