Corner Matching Refinement for Monocular Pose Estimation

Dinesh Gamage
dinesh.gamage@moansh.edu
Tom Drummond
tom.drummond@monash.edu
Monash University
Australia

Abstract

Many tasks in computer vision rely on accurate detection and matching of visual landmarks (e.g. image corners) between two images. In particular, for the calculation of epipolar geometry from a minimal set of five correspondences the spatial accuracy of matched landmarks is critical because the result is very sensitive to errors.

The most common way of improving the accuracy is to calculate a sub-pixel location independently for each landmark in the hope that this reduces the re-projection error of the point in space to which they refer. This paper presents a method for refining the coordinates of *correspondences* directly. Thus given some coordinates in the first image, our goal is to maximise the accuracy of the estimate of the coordinates in second image corresponding to the same real world point without being too concerned about which real world point is being matched.

We show how this can be achieved as a frequency domain optimisation between two image patches to refine the correspondence by estimating affine parameters. We select the correct frequency range for optimisation by identifying a direct relationship between the Gabor phase based approach and the frequency response of a patch. Further, we show how parametric estimation can be made accurate by operating in the frequency domain.

Finally, we present experiments which demonstrate the accuracy of this approach, its robustness to changes in scale and orientation and its superior performance by comparison to other sub-pixel methods.

1 Introduction

Accuracy of 3D structure calculation from an image sequence depends on accurately computing the motion of the camera. This in turn requires reliable feature extraction and matching. A particular problem that drives this work is that of calculating the essential matrix that describes the epipolar geometry of two images for which the internal camera parameters are known. This can be done from a minimal set of five correspondences between the two images using a polynomial solving algorithm [18] which can generate up to 10 Essential matrices, or by iterative optimisation of the residual error. In either case, the hypothesis generated from five matches is extremely sensitive to the accuracy with which the matches are extracted from images.

To improve the accuracy of the generated hypothesis, descriptor based feature matching which works at pixel level may not be adequate and sub-pixel level information may

http://dx.doi.org/10.5244/C.26.38

be needed. A number of applications already use something better than just pixel sampled information. Sub-pixel methods have been used extensively for stereo matching [19][15]. But most of these techniques are based on the assumption that the 2-D image motion, resulting from 3-D camera motion can be described using a simple translation model[6]. Widely used sub-pixel methods can be categorized as interpolation based methods (correlation interpolation, intensity interpolation and geometric methods)[1], phase correlation methods and differential methods (optical flow and parameter optimisation)[22].

Phase correlation has worked well for sub-pixel registration. But conventional phase correlation techniques fail when the matching window under consideration becomes small. Recent work has shown the necessity to fit a function to the phase correlation measurement in order to get satisfactory results [20]. Though these methods can be extended to sub-pixel patch matching/refinement, their applicability is limited to simple translations where any affine transformation needs to be rectified separately at a prior stage. Differential methods use a constraint equation under intensity conservation assumption [11][12][17] or handle the problem as an optimisation over a set of parameters[2], which works well under local patch deformations. Therefore in recent years, a considerable attention has been given to more complex motion models based on parameter estimation [4]. Such methods based on hierarchical or multi-resolution approaches have a limited applicability in time critical applications. But, because of the noise sensitivity and the better convergence rate, later parametric motion model has been extended in to the frequency domain [13, 14]. Such frequency domain approaches have shown better performance and noise tolerance compared to spatial domain methods. These methods use the shift invariance of the magnitude spectra to first estimate four non-translational affine parameters. The translation is then estimated using phase correlation between affine rectified images.

But in the frequency domain, the phase contains much information compared to the magnitude [7, 10] and shows a better robustness to noise [8]. Without discarding the phase, simultaneous optimisation of all six parameters yields better results. We parameterise the signal using the six parameter affine model with an additional parameter to compensate for energy changes of the signal. In pose estimation, the illumination between two consecutive frames won't change significantly. So the effect of the seventh parameter is trivial for our application. By optimizing in the frequency domain, it is possible to achieve improved results and a fast convergence rate. The fast convergence is a result of the multi-resolution nature that naturally arises with such an approach as we explain later. The following summarises our approach and contributions:

- For sub-pixel refinement, we represent local affine transformations of an image patch in the frequency domain and optimise over a set of parameters simultaneously, using both magnitude and the phase of the signal.
- We model local transformations of a projectively transformed image pair by an affine transformation, selecting a 32x32 patch around two corresponding corners and then try to refine the second corner position by affine warping frequency spectrum of the first patch and changing the phase of the second.
- In order to further increase the accuracy, we re-sample the second patch using the estimated translation.
- We derive a relationship between the Gabor filter phase difference and the frequency representation of a Gaussian weighted image patch and use this to select the effective frequency range for the optimisation.

- Using several re-sampling stages in the frequency and the spatial domains we get better sub-pixel accuracies (down to 0.1 pixels under moderate affine transformations) and a better convergence.

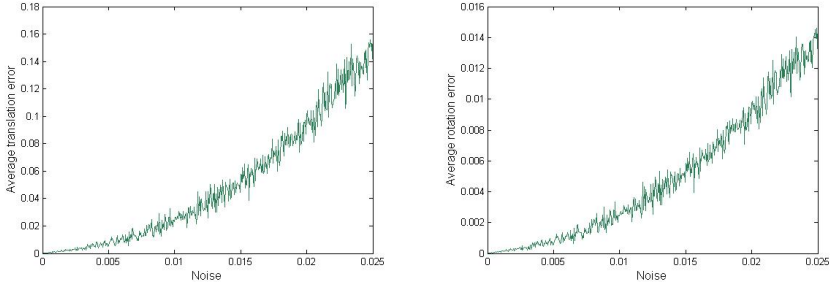
- These sub-pixel refined correspondences are then used to get a more stable and an accurate pose estimate.

The remainder of this paper is structured as follows. Section 1.1 briefly analyses the sensitivity of the camera pose to pixel noise and justify the necessity of a pixel-refinement stage. Section 2 describes the frequency domain affine parameterisation method with a criterion for selecting the correct frequency range. In section 3 we demonstrate the effectiveness of the new method for sub-pixel refinement with experimental results and conclude with a brief discussion.

1.1 Monocular pose estimation

The sensitivity of Essential matrix calculation to correspondence data is partly a consequence of nonlinear error propagation with depth which leads to a deviation from the desirable Gaussian uncertainty representation. Different parametrisation techniques have been proposed to reduce this non linearity to get better results [16]. Though these methods are capable of making the pose estimation less sensitive to pixel noise, it still remains as a major source that corrupts the final estimate [5].

Figure 1 shows the results of a simulated experiment that illustrates this point. In this experiment, 100 3D landmarks were randomly generated around a specified average depth (20) from the first camera. These points were then projected into a second camera with a translation of 1 unit away from the first, with a random rotation. Then an isotropic measurement noise was added to these projected locations. All corresponding points from two views were then used to compute the least squares approximation to the essential matrix from which the translation and the rotation were recovered to compare the result with the ground truth. It can be seen that the absolute error for both translation and rotation increases rapidly when the average noise level is increased.



(a) Translation error vs noise (b) Rotation error vs noise

Figure 1: Pose estimation absolute error (estimated from two artificially projected camera frames) vs the maximum noise magnitude

2 Estimation of the affine transformation

In this section we present the affine parameter model in the frequency domain. We make use of the affine theorem in the frequency domain [3]. Given two image patches $I_0(\bar{x})$ and $I_1(\bar{x})$ surrounding two corresponding corners, which are related by an affine coordinate transformation $I_1(\bar{x}) = I_0\left[A^{-1}\left(\bar{x}-\bar{b}\right)\right]$, where A and b are the four non translational affine parameters and two translational parameters respectively, their 2-D Fourier transforms are related by:

$$\hat{I}_1(\bar{u}) = |det\,(A)|\, e^{-j\bar{u}.\bar{b}}\hat{I}_0\left(A^T\bar{u}\right) \tag{1}$$

The shift invariance property of the magnitude spectra of equation 1 enables the estimation of $\bar{b}$ to be separated from the estimation of A [12]. But discarding phase is a huge waste as phase carries a lot of information in the frequency domain, which can be used to get more stable and fast estimations by simultaneously optimising all six parameters.

Here we use the six parameter affine model with an additional parameter. The seventh parameter compensates for energy changes caused by different local illumination conditions. If we select $\bar{\beta} = \{\beta_1 \ldots \beta_7\}$ to be the parameter set and absorb the $|det\,(A)|$ of the equation 1 into β_7 we have:

$$\beta_7\hat{I}_1(\bar{u}) = e^{-j\bar{u}.\bar{b}}\hat{I}_0\left(A^T\bar{u}\right) \qquad \text{where} \quad A = \begin{pmatrix}\beta_1 & \beta_2\\ \beta_3 & \beta_4\end{pmatrix} \quad \text{and} \quad \bar{b} = \begin{pmatrix}\beta_5\\ \beta_6\end{pmatrix} \tag{2}$$

$$\text{so} \quad \beta_7 e^{j\bar{u}.\bar{b}}\hat{I}_1(\bar{u}) = \hat{I}_0\left(A^T\bar{u}\right) \tag{3}$$

Thus the error r, for a frequency $\bar{u}$ can be written as,

$$r\left(\bar{u},\bar{\beta}\right) = \beta_7 e^{j\bar{u}.\bar{b}}\hat{I}_1(\bar{u}) - \hat{I}_0\left(A^T\bar{u}\right) \tag{4}$$

The above equation enables us to model the affine transformation as a phase change of $\hat{I}_1$ and a warp of $\hat{I}_0$ with respect to matrix A. The Jacobian J_i of partial derivatives of r with respect to β_i can then be computed:

$$[J_1, J_2, J_3, J_4, J_5, J_6, J_7] = \left[-\frac{\partial I_0}{\partial u}u, -\frac{\partial I_0}{\partial u}v, -\frac{\partial I_0}{\partial v}u, -\frac{\partial I_0}{\partial v}v, -\beta_7\tilde{I}_1 u, -\beta_7\tilde{I}_1 v, \tilde{I}_1\right] \tag{5}$$

Given a set of frequencies $\{u_j\}$, the errors $r(u_j)$ and the Jacobian J_{ij} can be used to obtain the parameters $\bar{\beta}$ that minimise $E = \sum_j \|r(u_j)\|^2$ using Gauss-Newton.

2.1 Iterative refinement

After initializing the set of parameters by setting A to be the identity and $\bar{b}$ to a zero vector, we use Gauss-Newton method to warp the frequency patch $\hat{I}_0$ with respect to the first four parameters $\beta_1 \ldots \beta_4$, and phase shift the patch $\hat{I}_1$ with the remaining two parameters β_5 and β_6. Warping is done by sub-sampling the original frequency patch using bilinear interpolation. After optimizing for two or three iterations in the frequency domain, we extract the parameters β_5 and β_6, which correspond to a translation in the spatial domain in x and y directions respectively. Then these two parameters are used to re-sample the second patch (patch I_1) in the spatial domain at the new refined position using spatial sub-sampling. The Fourier transform of this patch is then used to re-estimate a new set of affine parameters. This routine is continued until sufficient accuracy is achieved. According to experimental results, two spatial sampling steps are usually sufficient to reduce the average pixel error down to 0.1 pixels.

2.2 Aliasing and the DC response

There are two issues that need to be addressed in order for this to work in practice. FFT requires a periodic signal. So each patch has to be compensated for edge effects at the border. Secondly, the presence of a large DC component in the signal corrupts low frequency components of the signal in the frequency domain (those where $\|u\|$ is small). To remove edge effects from the image patches, we multiply it by a Gaussian weighting window ($G(x,y)$) centered at the detected landmark before taking the Fourier transform. Before doing that, the DC component of the patch which appears as a large spike at $u = 0$ in the frequency domain can be removed by subtracting the average, which gives a new patch. After alleviating both of these effects we get a new patch $I'(x,y)$ defined as:

$$I'(x,y) = G(x,y)\left(I(x,y) - \frac{\sum_{x,y} G(x,y)I(x,y)}{\sum_{x,y} G(x,y)}\right) \tag{6}$$

This gives a patch with a 0 DC coefficient. The frequency response of the Gaussian multiplied patch, $\mathscr{F}[I']$ has a direct relationship with the Gabor filter with an identical Gaussian support. We use this relationship to select the useful frequency range (in order to eliminate possible aliasing effects) for the optimisation in a multi resolution manner.

Multiplying the patch by a Gaussian in the spatial domain is equivalent to a convolution by the Fourier transform of the Gaussian in the frequency domain. So each frequency point of the signal $\mathscr{F}[I']$ represents the average of the Fourier transform of the original signal around that point. This resembles the effect of a Gabor filter. Fourier transform of a Gabor filter is a shifted Gaussian in the frequency domain. This makes frequency points of the Fourier transform of I' to be responses of the original patch (with applied DC offset) to a set complex Gabor filters. This interpretation can be used to select the frequency range for the optimization.

Because the phase of a particular Gabor filter response changes linearly under spatial translations of the signal, it can be used to estimate spatial disparity of two instances of a the same signal with a relative shift [9]. This phase disparity is useful only if the displacement is smaller than a half a wavelength of the tuning frequency [9] of the Gabor filter, i.e the domain of convergence for the phase is $\pm\pi$. This imposes an upper frequency limit for the useful frequency range. If we assume a maximum displacement of d pixels for a 1-D signal this criteria suggests a frequency f such that $f \leq 1/2d$. In 2-D this requirement can be met by limiting the useful frequency range radially to a maximum of $1/2d$ radius. After estimating the translation (and other parameters) using small frequencies for large displacements finer refinements can be done gradually by increasing the radius, incorporating higher frequency responses to the optimization.

Though subtracting the DC component spatially as in equation 6 can mostly reduce its effect, for better results we have to impose a lower frequency limit as well. We select the minimum frequency using the one octave bandwidth criteria which has been suggested in the literature [8] for Gabor filter based disparity estimations. The one octave bandwidth in the frequency domain for the Gabor filter shows the spatial support to be:

$$\sigma = \frac{1}{2\pi f}\left(\frac{2^\alpha + 1}{2^\alpha - 1}\right) \tag{7}$$

If we select above frequency f keeping the spatial support σ a constant, in order to

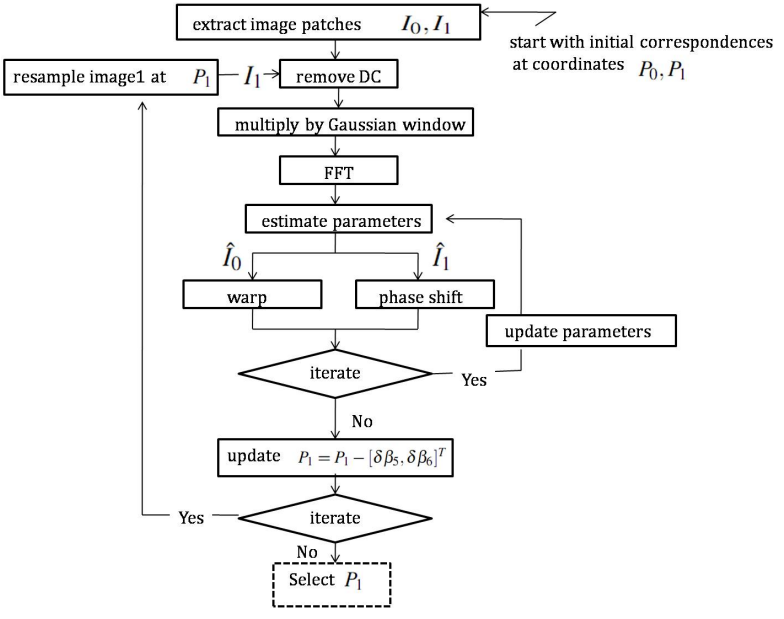


Figure 2: Iterative optimization based on sub-sampling in the frequency and the spatial domains

eliminate any DC distortion the minimum frequency should be:

$$f \geq \frac{1}{2\pi\sigma}\left(\frac{2^\alpha + 1}{2^\alpha - 1}\right) \tag{8}$$

Combining the minimum and the maximum criteria for frequency selection gives:

$$\frac{1}{2d} \geq f \geq \frac{1}{2\pi\sigma}\left(\frac{2^\alpha + 1}{2^\alpha - 1}\right) \tag{9}$$

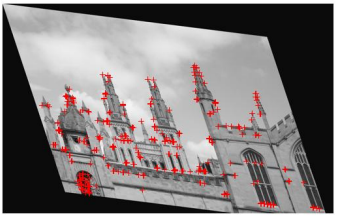
At the end of each iteration we can expect the displacement d to reduce, increasing the useful frequency range. Higher frequencies carry finer details about the translation, which improves the final solution. This naturally enables a multi-resolution framework for refinement without any additional computations.

Figure 2 summarises the steps we use to sub-pixel refine a target corner position with respect to the given reference.

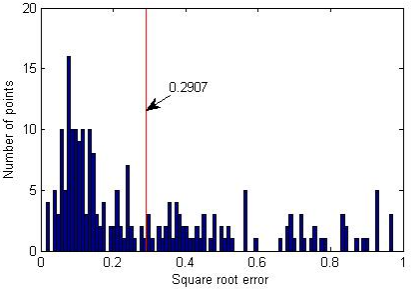
3 Results

In this section we results of applying the proposed refinement method to refine corner correspondences and compare it to a baseline spatial refinement method. The spatial method performs an iterative Gauss-Newton optimisation over a set of seven affine parameters to minimise the sum of the sqaured differences between I_0 and the affine transformation applied to I_1, sampled with bilinear interpolation: $\sum_{\bar{x}} I_1(\bar{x}) - I_0\left[A^{-1}\left(\bar{x} - \bar{b}\right)\right]$.

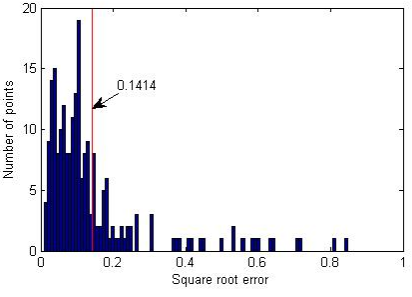
(a) Original image



(b) Transformed image



(c) Error distribution after spatial domain refinement



(d) Error distribution after frequency domain refinement

Figure 3: Reference and transformed image pairs error distributions.

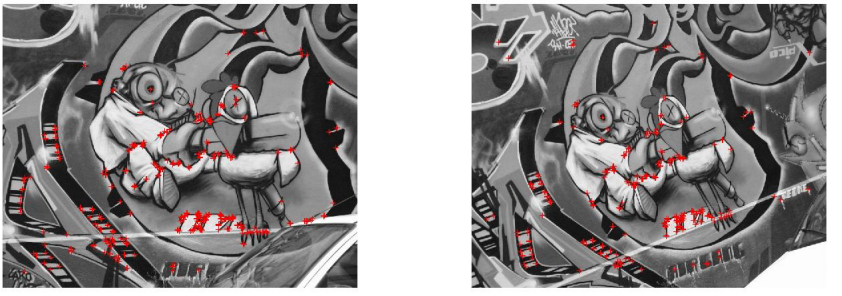
Here we refine point correspondences between two synthetically generated affine transformations, homographies obtained from two real planar views and for epipolar geometry calculation between two frames of a real 3D scene.

3.1 Synthetically generated transformations

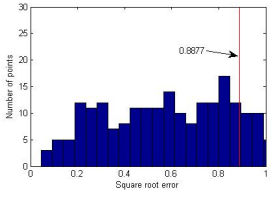
We first performed an experiment to demonstrate the improvement that can be achieved by two refinement methods in a situation where the ground truth transformation between two images was known. Synthetic data was generated first by transforming a reference image using a known affine transformation with bilinear interpolation. To remove interpolation artifacts, both images were down scaled by a factor of two. Then FAST features were extracted from the first image and projected into the second, using the known transformation. Projected corners were then rounded off to the nearest integer pixel. The raw pixel errors were then calculated as the distance between the ground truth and the rounded off positions. Figure 3 shows typical performance. With this setup the average unrefined pixel error is 0.4102. The spatial domain refinement method reduced the average error to 0.2907, while the frequency based method reduced it further to 0.1414.

3.2 Real image homographies

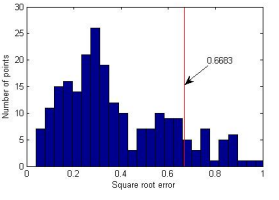
Here, we applied the method to a pair of images from the Graffiti database shown in Figure 4(a). FAST features were extracted from both images and matched using HIPS [7]. Inliers were then found by applying RANSAC to these raw matches and fixed for all stages of the



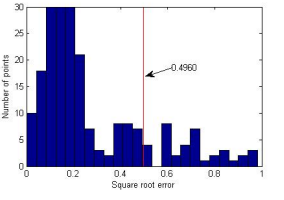
(a) Graffiti image pair



(b) Unrefined



(c) Spatial



(d) Frequency

Figure 4: Residual error distribution for the first image pair in Graffiti database

experiment. The inlier matches were then refined using both spatial and frequency methods. Homographies were then computed using the raw, spatial and frequency refined matches minimising $\sum_i r_i^2$ where $r_i = \sqrt{\|x_i - Hy_i\|^2 + \|y_i - H^{-1}x_i\|^2}/2$ is the residual error of a match and x_i is the homogeneous image coordinates of a FAST feature in image 1 and y_i is the homogeneous image coordinates of the refined location of its match in image 2. Figure 4 shows the distribution of the residual errors r_i for each of the three schemes (raw, spatial and frequency) applied to this image pair. The average raw error was 0.8877, the spatial refinement method reduced this to 0.6683 and the frequency refinement method reduced it to 0.4960.

3.3 Pose estimation

As discussed in Section 1.1, pose estimation from an Essential matrix is extremely sensitive to matching errors, and benefits from sub-pixel refinement. For a set of image pairs we

		Residual error in pixels for image pairs		
		P1	P2	P3
HIPS	Raw	0.447879	0.571511	0.535419
	Refined	0.360749	0.384233	0.373508
Sub-pixel SIFT		0.541322	0.435442	0.690141

Table 1: Comparison of Frequency-based refinement with sub-pixel features from SIFT

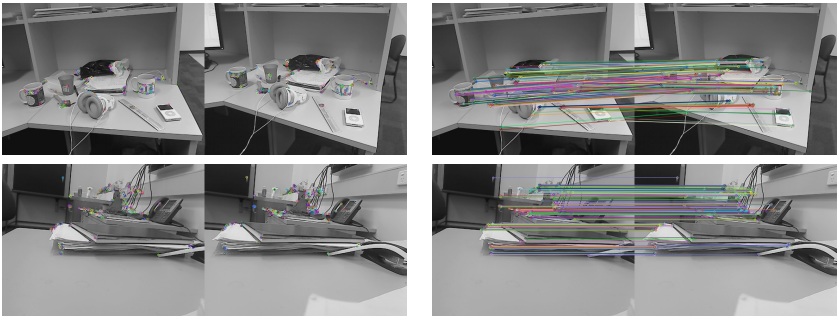


Figure 5: Two image pairs used for pose estimation

estimated Essential matrices using the interative five-point pose algorithm for FAST corners matches with HIPS. The symmetric residual point to line error was then calculated for each correspondence. These matches were then refined using frequency domain method and the Essential matrix was re-estimated. The residual errors were then recalculated. We compare our results with sub-pixel SIFT results. Table 1 summarises those errors. Figure 5 shows image pairs used for pose estimation.

4 Discussion and conclusion

In this paper we introduced an affine parametric model for match refinement that operates in the frequency domain. By making maximum use of phase information, we are able to obtain an accurate parameter estimation, in particular of the translation, that can be applied to obtain the best match. Most importantly we have shown the ability of the newly proposed method to refine correspondences in a coarse-to-fine multi-resolution manner in the Fourier domain.

Experimental results establish the effectiveness of the proposed method for modeling local patch deformations which can be used for sub-pixel refinement. Such locally refined corners can be then used to estimate the global monocular pose with improved accuracy. As a post processing step, after a less accurate but fast descriptor based feature matching stage our method can be used for efficient sparse match refinement.

However, due to the fixed size of the Gaussian weighting function, we have found that the refinement accuracy is sensitive to scale changes of more than $20-30\%$. If the image pair contain scale changes larger than this, some image pyramids scheme would be necessary. Further, if the translation is large compared to the minimum half wave-length of the selected frequency band, the solution degenerates as the Hessian matrix in Gauss-Newton algorithm becomes ill conditioned. Thus the accuracy of the final result depends on the coarse-to-fine frequency tuning of the optimization. Fortunately, in practice, the feature detection and matching methods used in this paper give correspondences that are well within the convergence band of our algorithm.

References

[1] C.A. Berenstein, L.N. Kanal, D. Lavine, and E.C. Olson. A geometric approach to subpixel registration accuracy. *Computer Vision, Graphics, and Image Processing*, 40 (3):334–360, 1987.

[2] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *Computer VisionÛECCV'92*, pages 237–252. Springer, 1992.

[3] RN Bracewell, K.Y. Chang, AK Jha, and Y.H. Wang. Affine theorem for two-dimensional fourier transform. *Electronics Letters*, 29(3):304, 1993.

[4] M. Campani and A. Verri. Motion analysis from first-order properties of optical flow. *CVGIP: Image Understanding*, 56(1):90–107, 1992.

[5] J. Civera, A.J. Davison, and J.M.M. Montiel. Inverse depth to depth conversion for monocular slam. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 2778–2783. IEEE, 2007.

[6] A. Donate, X. Liu, and EG Collins. Efficient path-based stereo matching with subpixel accuracy. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 41(1):183–195, 2011.

[7] G. Duan and R.A. Morris. The importance of phase in the spectra of digital type. *Electronic Publishing*, 2(1):47–60, 1989.

[8] D.J. Fleet and A.D. Jepson. Stability of phase information. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(12):1253–1268, 1993.

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

[10] M. Hayes, J. Lim, and A. Oppenheim. Signal reconstruction from phase or magnitude. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(6):672–680, 1980.

[11] D.J. Heeger. Model for the extraction of image flow. *JOSA A*, 4(8):1455–1471, 1987.

[12] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial intelligence*, 17 (1-3):185–203, 1981.

[13] T.I. Hsu, AD Calway, and R. Wilson. Texture analysis using the multiresolution fourier transform. In *8th Scandinavian Conference on Image Analysis*, pages 823–830, 1993.

[14] SA Kruger and AD Calway. A multiresolution frequency domain method for estimating affine motion parameters. In *Image Processing, 1996. Proceedings., International Conference on*, volume 1, pages 113–116. IEEE, 1996.

[15] L. Matthies, T. Kanade, and R. Szeliski. Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3(3):209–238, 1989.

[16] JMM Montiel, J. Civera, and A.J. Davison. Unified inverse depth parametrization for monocular slam. *analysis*, 9:1, 2006.

[17] H.H. Nagel. On the estimation of optical flow: Relations between different approaches and some new results. *Artificial Intelligence*, 33(3):299–324, 1987.

[18] D. Nistér. An efficient solution to the five-point relative pose problem. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(6):756–770, 2004.

[19] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002.

[20] T. Shibahara, T. Aoki, H. Nakajima, and K. Kobayashi. A sub-pixel stereo correspondence technique based on 1d phase-only correlation. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, volume 5, pages V–221. IEEE, 2007.

[21] S. Taylor and T. Drummond. Multiple target localisation at over 100 fps. 2009.

[22] Q. Tian and M.N. Huhns. Algorithms for subpixel registration. *Computer Vision, Graphics, and Image Processing*, 35(2):220–233, 1986.