

# Fast, Unconstrained Camera Motion Estimation from Stereo without Tracking and Robust Statistics

Heiko Hirschmüller, Peter R. Innocent and Jon M. Garibaldi  
Centre for Computational Intelligence,  
De Montfort University, Leicester, LE1 9BH, UK,  
hhm@dmu.ac.uk, pri@dmu.ac.uk, jong@dmu.ac.uk.

## Abstract

Camera motion estimation is useful for a range of applications. Usually, feature tracking is performed through the sequence of images to determine correspondences. Furthermore, robust statistical techniques are normally used to handle large number of outliers in correspondences. This paper proposes a new method that avoids both. Motion is calculated between two consecutive stereo images without any pre-knowledge or prediction about feature location or the possibly large camera movement. This permits a lower frame rate and almost arbitrary movements. Euclidean constraints are used to incrementally select inliers from a set of initial correspondences, instead of using robust statistics that has to handle all inliers and outliers together. These constraints are so strong that the set of initial correspondences can contain several times more outliers than inliers. Experiments on a worst-case stereo sequence show that the method is robust, accurate and can be used in real-time.

## 1 Introduction

### 1.1 Camera Motion Estimation

This paper describes a novel method that estimates motion of a calibrated stereo camera, which moves unconstrained in a static scene. The long-term goal is to use motion estimation to integrate stereo data incrementally in real-time into a global model of a scene as the camera moves through it.

The stereo camera will be mounted on a tele-operated mobile robot for the current application. However, the method does not use any sensory information from the robot itself as an initial estimate of camera motion. It relies only on the dense output of a real-time correlation-based stereo vision system, which has been described earlier [1].

The computer hardware on-board the robot is expected to be limited. Thus, the method has to be fast so that it works additionally to the stereo system in real-time. Additionally, the method must robustly handle large changes in images, due to low frame rates.

### 1.2 Review of Existing Methods

Camera motion estimation is essential for many applications, e.g. structure from motion (SFM). General SFM techniques can reconstruct a scene, which is captured from different viewpoints using a single, possibly un-calibrated camera [2]. Techniques for calibrated stereo cameras

can take advantage of 3-D measurements at every viewpoint to increase reliability and avoid degenerative cases [3, 4, 5, 6, 7].

Calculating motion can be done from seven correspondences in the image planes by recovering the fundamental matrix [8]. However, certain point configurations can degrade the accuracy or lead to a complete failure, e.g. if all points are co-planar in the scene. On the other hand, three 3-D point correspondences, which are not co-linear permit the calculation of rigid motion.

Another important issue is whether the method processes all images off-line together or incrementally, as the camera moves through the real world. The later ones are suitable for a real-time use on mobile robots [9, 3, 4, 5].

The techniques are usually based on features, whose correspondences are determined in images that are taken from different viewpoints. Most commonly, point features are used, because they are accurately to localise [2, 10, 9, 3, 4, 5]. Other methods use 3-D lines or planes, which are available through feature based stereo [7, 6].

Correspondences can for features be determined by correlation of the area around the feature [3, 4, 5, 2]. However, changes in projection, rotation and scale need to be accounted for. Certain invariants can help to select appropriate point features as well as to find correspondences [10]. Additionally, tracking of features through the image sequence allows to predict their location and thus to limit possible correspondences [3, 4, 5]. Features with richer descriptions may be matched directly by matching their parameters, possibly with taking even error characteristics of the capturing device into account [6, 7].

However, there is usually a substantial amount of outliers in correspondences, which can drive the result far away from the solution. This problem is normally tackled with a robust statistical technique, like weighted least squares [11], a random sampling approach (e.g. RANSAC) [3] or some other process that suppresses the effect of outliers iteratively [6, 4, 2, 5].

### 1.3 Proposed Constraint Satisfaction Method

The basic motivation of the new method is to exploit the increased information that a stereo camera offers over a single camera. The 3-D position of point features is determined through stereo vision. Distances between the 3-D positions of features remain obviously the same, regardless

of the position and orientation of the camera. Thus, an initial guess on the correspondence of two features can easily be validated by comparing their distance in both camera coordinate systems. If the distances differ, then at least one of the correspondences is definitely wrong.

It turns out that this constraint is so strong that it can detect almost all outliers, even if the amount of outliers is several times higher than the amount of inliers. Furthermore, all remaining outliers that pass the test by accident do not introduce high errors in the result, because they are enforced to be close to their correct position.

Section 2.1 provides an overview over the whole method and describes how the techniques are combined with existing ones. Section 2.2 and 2.3 discuss the novel ideas about establishing initial correspondences and the fast detection of outliers in detail. Results from experiments are discussed in section 3. Section 4 concludes the paper.

## 2 The Constraint Satisfaction Method

### 2.1 Overview

An overview over the whole method is shown in figure 1. The rectified left and right images are used to calculate a dense disparity image, using an earlier developed real-time stereo system [1]. Additionally, the left camera image is used to detect corners with the Harris corner detector [12]. This produces typically several hundred points. The Harris corner detector is known for its reliability to detect the same features again (i.e. no flickering of corners in consecutive images), which is important for this application.

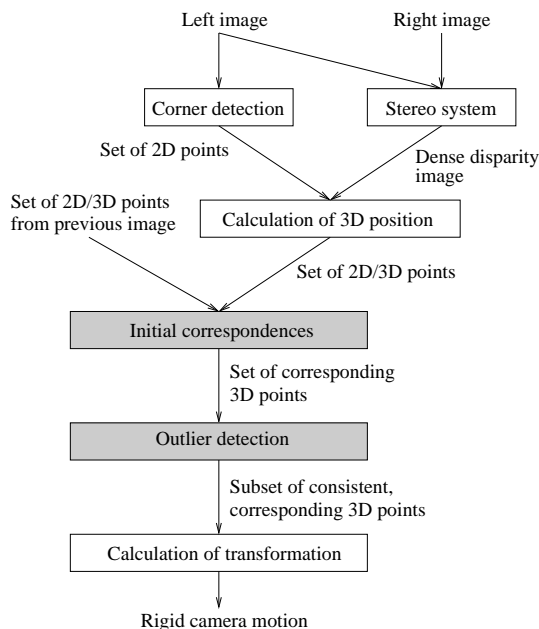


Figure 1: Overview, grey boxes represent novel parts.

Next, the 3-D position of each point is calculated, using the disparity image. The disparity image usually contains areas for which the disparity can not be determined (e.g. texture-less areas). Points that fall in or are very close to such an area are discarded. Furthermore, all points which are further away than a certain maximum distance (e.g.

10m) are not considered, due to an increased uncertainty in their position (see section 2.3.2).

The next step determines correspondences of points from the previous and current image. This is explained in detail in section 2.2. The result is expected to contain a high amount of outliers. The detection of outliers is described in detail in section 2.3. The result of outlier detection is a set of corresponding points, whose relative distances between each other are all maintained. If there are outliers left, then their influence is expected to be low.

Therefore, a direct least squares solution is sufficient to calculate the transformation between the sets of points [11]. The accuracy of the transformation is further increased by an iterative optimisation that is based on the true reconstruction characteristic of stereo vision [13].

This paper focuses on the determination of correspondences and outlier detection, since the remaining steps are straight forward once outliers are removed.

### 2.2 Initial Point Correspondences

#### 2.2.1 General Considerations

Two points in two different images correspond if they represent the same object in the real world. The correspondence between relevant points from consecutive left camera images is required. It is assumed that there will be a high amount of points, for which no correspondence exists, because some points are outside the field of view of one of the cameras. Furthermore, points can be occluded by objects. Finally, corner detection can fail to detect a corner in one image, but detects it in the other. It is not important to find all correspondences. Theoretically, three non-collinear points are sufficient to calculate motion. However, a higher amount is useful to increase accuracy, because the effect of noise can be reduced.

Many techniques perform a kind of tracking of feature points to predict their location in a new image. The proposed method is designed for a system with a low frame rate in respect to the anticipated motion. This results in highly differing images. Thus, motion might be too unpredictable to perform tracking. Therefore, no information about earlier images is used.

The easiest way to determine correspondences is by correlating the image area around a feature point with all possible places in the second image and choosing the place that matches best. However, large changes of the view-point projects the same area differently on the image plane. Furthermore, illumination might vary significantly. All of this results in high amounts of outliers in corresponding pixels within the compared area, which decreases the performance of correlation. A simple correlation, like the sum of absolute differences (SAD), will in many cases fail.

#### 2.2.2 Non-Parametric Correlation with Consistency Check

Experiments with stereo correlation methods showed several measures that can deal with outliers in the correlation calculation [1]. The non-parametric Rank and Census correlation can handle outliers well and they are insensitive in

changes of illumination [14], because they are based on the local ordering of intensities rather than the intensity values themselves.

The SAD5 measure that has been proposed in the study about stereo correlation [1] performs better than Rank and Census in case of stereo vision, but it is expected to perform worse for the current problem. Outliers in stereo vision appear usually at places where a correlation window overlaps a depth discontinuity. Hence, outliers are clustered together in a part of the window. SAD5 works by rejecting the part of the correlation window that is supposed to have a high amount of outliers. Rank and Census allow outliers to appear everywhere, which is expected to be the case when deformed areas are correlated as in the current application. However, the matching capability of Rank and Census is weaker than that of SAD.

Establishing initial correspondences is implemented by correlating all feature areas of one image against all feature areas of the other image, using Rank or Census with a big correlation window (e.g. 19x19 pixel). A consistency check then performs correlation in the other way around, by searching correspondences from the second to the first image. Only correspondences, which are found in both directions are retained.

The consistency check makes it likely that the correspondences are correct. However, outliers are still expected to occur and their influence can be very high, since there has been no restriction imposed on the parts of the image in which correspondences are sought (i.e. no tracking).

## 2.3 Detection of Outliers

### 2.3.1 Formal Basis of the Method

The initial correspondence step provides a set of corresponding points, whose 3-D position is known in respect to their camera coordinate system. The motion between two camera coordinate systems is described by a rotation  $R$  and a translation  $t$ . The relation between the points of the previous and current viewpoint ( $P_i$  and  $C_i$ , with corresponding points having the same index  $i$ ) can be written as,

$$C_i = RP_i + t \quad \text{with } i = 1 \dots n. \quad (1)$$

A property of rigid transformations in Euclidean space is that the relative distances between transformed points are maintained. Thus,

$$|P_i - P_k| = |C_i - C_k| \quad \text{with } i, k = 1 \dots n. \quad (2)$$

Another constraint can be derived if the rotation  $R$  is restricted. Every rotation in 3-D can be represented by a rotation axis  $r$  and a rotation angle  $\phi$ . If the vector  $P_i - P_k$  is orthogonal to  $r$ , then the angle between  $P_i - P_k$  and  $C_i - C_k$  is exactly  $\phi$ . If  $P_i - P_k$  is parallel to  $r$ , then the angle between  $P_i - P_k$  and  $C_i - C_k$  is always zero. In all other cases, the angle is in between 0 and  $\phi$ .

The rotation axis  $r$  is unknown as well as the angle  $\phi$ . However, imposing the restriction  $\phi < \theta$  results in the inequality,

$$(P_i - P_k)(C_i - C_k) > \cos \theta \quad \text{with } i, k = 1 \dots n. \quad (3)$$

For this study,  $\theta$  was set to  $\frac{\pi}{4}$ , which is a very low restriction on camera movement, since the camera may rotate by up to 45 degrees between two consecutive images. Still, this additional constraint is useful to identify outliers.

Basically, if the constraints (2) and (3) hold for two correspondences  $i$  and  $k$ , then both correspondences may be correct. However, if one of the constraints does not hold then at least one of the correspondences is definitely wrong.

A consistent subset of initial corresponding points can be constructed by selecting the largest number of points for which (2) and (3) holds for all combinations of  $i$  and  $k$ . The larger this subset is the more likely it is that it represents only inliers. Though, outlier can still exist. Certain point configurations can for example be constructed artificially, which are ambiguous, e.g. constraint (2) *alone* is ambiguous if four points are arranged in a square. However, such configurations are very rare in natural scenes, especially when the number of inliers is high enough.

There are two problems left. Firstly, constraint (2) is almost always wrong, for noisy values. The consistency of the distance needs to be determined by using the special error characteristics of stereo vision (see section 2.3.2). Secondly, naive implementations of above constraints could result in trying to solve a NP problem. Section 2.3.3 describes an algorithm that creates a consistent subset.

### 2.3.2 Error Characteristics of Stereo Vision

Stereo vision systems have the characteristic that errors propagate non-linear from the image plane into the 3-D position of a point, as shown in this section. This special characteristic needs to be taken into account for comparing distances properly.

Equation (4) describes the calculation of the 3-D coordinates  $X$ ,  $Y$  and  $Z$  from two corresponding points in the left and right image plane  $x_l, y_l$  and  $x_r, y_r$ <sup>1</sup> for the general case that  $y_l$  is not equal to  $y_r$  due to errors. The parameter  $f$  is the focal length and  $t$  the baseline of the stereo system.

It must be noted that this model is a simplification, because planar rectification does not restrict the focal length of both cameras in the direction of the base line, which is usually horizontal. Only the vertical focal length must be the same. However, the horizontal focal lengths have usually very similar values. In fact, they have to be very similar, so that correlation based stereo vision can work. Otherwise, two windows could not be correlated without taking different scale factors into account.

$$X = \frac{x_l t}{x_l - x_r} \quad Y = \frac{t(y_l + y_r)}{2(x_l - x_r)} \quad Z = \frac{f t}{x_l - x_r} \quad (4)$$

If the error  $\Delta e$  in the pixel positions in the image planes is modelled by independent Gaussian noise, then it results in the errors  $\Delta X$ ,  $\Delta Y$  and  $\Delta Z$  according to the rules of error propagation, i.e.

<sup>1</sup>Using pixel coordinates, which are centred in the principal point.

$$\Delta X = \Delta e \frac{Z}{ft} \sqrt{(t - X)^2 + X^2}, \quad (5)$$

$$\Delta Y = \Delta e \frac{Z}{ft} \sqrt{2Y^2 + \frac{1}{2}t^2}, \quad (6)$$

$$\Delta Z = \Delta e \frac{Z^2}{ft} \sqrt{2}. \quad (7)$$

Molton and Brady derived the same results from different considerations [3]. The distance  $L$  between two 3-D points is calculated by,

$$L = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2}. \quad (8)$$

It's propagated error  $\Delta L$  is calculated from the error in the image plane  $\Delta e$  by,

$$\Delta L = \frac{\Delta e}{Lft} \sqrt{Z_1^2(A + B + C) + Z_2^2(D + E + F)} \quad (9)$$

$$A = ((X_1 - X_2)(t - X_1) - (Y_1 - Y_2)Y_1 - (Z_1 - Z_2)Z_1)^2$$

$$B = ((X_1 - X_2)X_1 + (Y_1 - Y_2)Y_1 + (Z_1 - Z_2)Z_1)^2$$

$$C = \frac{1}{2}(t(Y_1 - Y_2))^2$$

$$D = ((X_1 - X_2)(t - X_2) - (Y_1 - Y_2)Y_2 - (Z_1 - Z_2)Z_2)^2$$

$$E = ((X_1 - X_2)X_2 + (Y_1 - Y_2)Y_2 + (Z_1 - Z_2)Z_2)^2$$

$$F = \frac{1}{2}(t(Y_1 - Y_2))^2.$$

Hence, the error  $\Delta L$  depends very much on the position of both 3-D points and their distance from the camera and must be calculated for every pair of points individually.

Finally, equation (10) can be used to check if two distances  $L_1$  and  $L_2$  with their errors  $\Delta L_1$  and  $\Delta L_2$  are consistent, i.e. if they are likely to represent the same distance. This replaces equation (2). The error  $\Delta e$  can be set to a small value, i.e. 0.2 pixels has been used for this paper.

$$|L_1 - L_2| \leq 3\sqrt{\Delta L_1^2 + \Delta L_2^2} \quad (10)$$

### 2.3.3 Handling Complexity

Determination of a large consistent subset of a set of  $n$  corresponding points is performed by using a  $n \times n$  matrix  $m$  as shown in figure 2. The elements of  $m$  denote if the corresponding points  $P_i, C_i$  and  $P_k, C_k$  passed the test (10) and (3) with 1 or not, with 0. The goal is to find the largest number of points, so that  $m_{ik} = 1$  (i.e.  $m_{ik} = m_{ki}$ ) for all combinations of points  $i, k$  in the set.

It seems that finding the largest number of points would involve to test all combinations of  $n$  points, which would require  $O(2^n)$  steps (i.e. it seems to be a NP problem). However, a good approximation is to find not the largest, but a large number of consistent points. This can be done

by first selecting the point  $i$  with the largest number of consistencies (i.e. the column  $i$  with the largest number of 1-elements). Further points  $k$  are incrementally added, so that  $m_{ik} = 1$  for all previous points  $i$  and as much as possible further consistencies are maintained. This is repeated until there are no further points to add.

	$P_1, C_1$	$P_2, C_2$	...	$P_n, C_n$	
$P_1, C_1$	1	0	...	0	← Inconsistent
$P_2, C_2$	0	1		1	← Consistent
...	...				
$P_n, C_n$	0	1		1	

Figure 2: Matrix  $m$  of point consistencies.

The time for creating matrix  $m$  is  $O(n^2)$ , while the time to create a consistent subset of correspondences is  $O(n_c n^2)$ , with  $n_c$  as the number of resulting consistent, corresponding points. It should be noted that  $n$  is usually a small number (e.g. 50) and that the operations after creating matrix  $m$  are very simple.

## 3 Results

An example sequence that consists out of 18 stereo images has been captured with the real-time stereo system. The camera has basically been moved in a 360 degree circle. Additionally, forward, backward and sideways movements have been introduced as well. A rotation around the optical axis has been avoided, since it is not normal for robotics applications. Typically, only half of an image overlaps with the previous one. Only a low resolution (i.e.  $320 \times 240$  pixel) is used and the lighting differs significantly. Thus, the sequence is considered to be a worst-case scenario. The images number 13 and 14, shown in figure 3 are a good example. The black lines represent corresponding points, found by the algorithm. All 17 correspondences were manually confirmed to be correct.

The last image overlaps with the first one of the sequence, which makes it possible to determine the overall error of motion estimation. Figure 4 shows images 18 and 1. The motion between both images contains a slight rotation around the optical axis as well. Nevertheless, there is only 1 outlier among 31 correspondences and this outlier is only a few pixels away from its correct position.

Images 7 and 8 are among the most problematic ones in the sequence. Some corresponding features were found on the structures of the wood or floor. It has been a problem to determine the correctness of the algorithmic choices as a human. 3 outliers have been found among only 8 correspondences. Nevertheless, the outliers are still near their correct position, so that the resulting motion is supposed to be still usable as a rough estimate.

Generally, around 300 corners are determined in each image. Around 140 of them are usable (e.g. their disparity is valid). The number of initial correspondences is lower due to the consistency check in correlation. The left diagram in figure 6 gives detailed results of the sequence using



Figure 3: Images 13 and 14 of the sequence (i.e. consecutive left stereo images). Black lines show consistent correspondences.

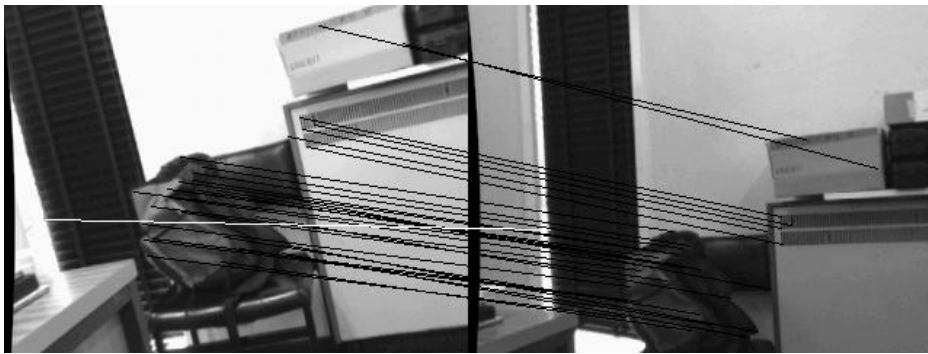


Figure 4: Images 18 and 1 of the sequence, i.e. last and first image. The white line shows a remaining outlier.



Figure 5: Images 7 and 8 of the sequence. There are three remaining outliers shown in white.

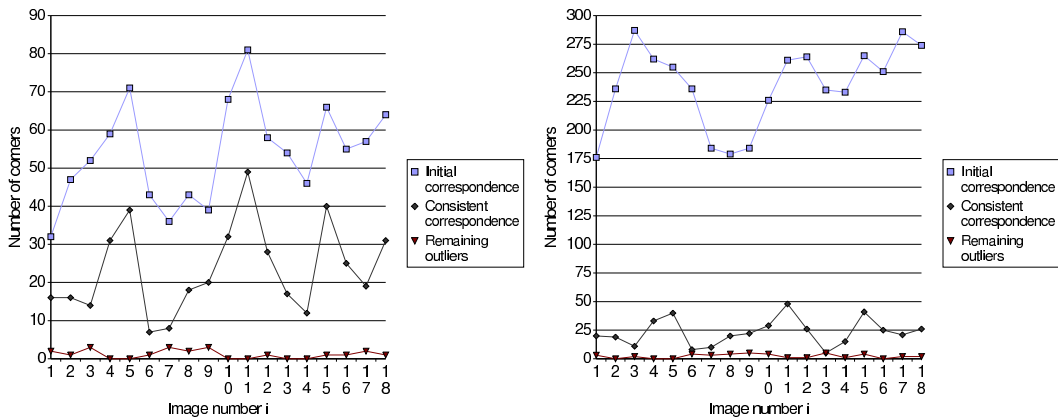


Figure 6: The number of initial correspondences, consistent correspondences and remaining outliers between image  $i$  and  $i+1$ . The normal small initial correspondence set is used for the graph on the left. The right graph shows results with a large correspondence set that contains much more outliers. Nevertheless, outliers are still detected robustly.

Census as a correlation measure. The consistent correspondences after outlier detection contain sometimes manually found outliers. However, all of them are near their correct position.

The amount of consistent correspondences in images 7, 8 and 9 is low and there is a substantial amount of outliers left. Results from motion calculation show a several times higher error (i.e. mean distance between corresponding points after transformation) than for all other images.

In spite of these problems, the orientation of the camera after closing the 360 degree cycle is only by 9 degrees vertically, 7 degrees horizontally and 2 degrees around the optical axis incorrect. The position of the camera after the cycle is obviously very sensitive to rotational errors. The camera travelled a distance of 3.69m and is in the end 0.01m away from its initial position. It is assumed, that images 7, 8 and 9 are mainly responsible for these errors, as explained above.

Finally, the diagram on the right of figure 6 demonstrates the power of the method to discriminate between inliers and outliers in a set that contains much more outliers. The consistency check has been removed. Instead all correspondences found by searching corresponding points from the first to the second and from the second to the first image are returned together. Thus, the set is around 4 times bigger with a much higher number of outliers. Nevertheless, the resulting sets of consistent correspondences are almost the same. This demonstrates the discriminative power of the method impressively.

The method has been implemented in C and some parts in MMX Assembler. The speed on the discussed sequence has been measured on an Athlon with 1.2GHz under Linux. Corner detection takes typically around 17ms per image. Initial correspondences are established in another 17ms using Rank as correlation measure, while outlier detection takes only around 1ms. The transformation is calculated in 3ms, with a least squares calculation followed by an iterative optimisation [11, 13].

## 4 Conclusion

A novel camera motion estimation method has been presented, which does not make use of feature tracking and robust statistics as most other methods. It has been shown that the method works with differences in lighting and large, almost arbitrary movements of the camera, resulting in an overlap of only half of an image. The method shows robustness and is fast. Its power to detect almost all outliers in a set containing several times more outliers than inliers has been demonstrated. Remaining outliers are enforced to be close to their correct position so that they do not affect motion estimation very much. Thus, motion calculation delivers even with very much differing images a good estimate.

Future plans include to improve accuracy by trying to find more correspondences after motion calculation and calculating motion again with a bigger set, thus reducing the effect of errors further. Furthermore, it is planned to integrate the stereo data into a global model using the camera

motion information. Not only mobile robot applications can finally benefit from this automatically generated global 3-D model.

## Acknowledgements

We would like to thank QinetiQ for their financial support.

## References

- [1] H. Hirschmüller, P. R. Innocent, and J. M. Garibaldi, "Real-time correlation-based stereo vision with reduced border errors," *International Journal of Computer Vision*, vol. 47, pp. 229–246, April-June 2002.
- [2] M. Pollefeys, R. Koch, M. Vergauwen, and L. V. Gool, "Flexible 3d acquisition with a monocular camera," in *Proceedings of the IEEE International Conference on Robotics and Automation, Vol. 4*, pp. 2771–2776, IEEE, May 1998.
- [3] N. Molton and M. Brady, "Practical structure and motion from stereo when motion is unconstrained," *International Journal of Computer Vision*, vol. 39, pp. 5–23, August 2000.
- [4] A. Mallat, S. Lacroix, and L. Gallo, "Position estimation in outdoor environments using pixel tracking and stereo vision," in *Proceedings of the IEEE International Conference on Robotics and Automation, Vol. 4*, pp. 3519–3524, IEEE, April 2000.
- [5] P. Saedi, P. Lawrence, and D. Lowe, "3d motion tracking of a mobile robot in a natural environment," in *Proceedings of the IEEE International Conference on Robotics and Automation, Vol. 2*, pp. 1682–1687, IEEE, April 2000.
- [6] N. Ayache and O. Faugeras, "Maintaining representations of the environment of a mobile robot," *IEEE Transactions on Robotics and Automation*, vol. 5, pp. 804–819, December 1989.
- [7] Z. Zhang and O. Faugeras, "A 3d world model builder with a mobile robot," *International Journal of Robotics Research*, vol. 11, pp. 269–285, August 1992.
- [8] O. Faugeras and Q.-T. Luong, *The Geometry of Multiple Images*. MIT Press, 2001.
- [9] A. Chiuso, P. Favaro, H. Jin, and S. Soatto, "3-d motion from 2-d motion causally integrated over time," in *Proceedings of the 6th European Conference on Computer Vision*, (Dublin), pp. 734–750, 2000.
- [10] M. Knapek, R. S. Oropeza, and D. J. Kriegman, "Selecting promising landmarks," in *Proceedings of the IEEE International Conference on Robotics and Automation, Vol. 4*, pp. 3771–3777, IEEE, April 2000.
- [11] R. Haralick, H. Joo, C.-N. Lee, X. Zhuang, V. G. Vaidya, and M. B. Kim, "Pose estimation from corresponding point data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, pp. 1426–1446, November, December 1989.
- [12] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the 4th Alvey Vision Conference*, pp. 147–151, 1988.
- [13] L. Matthies and S. A. Shafer, "Error modeling in stereo navigation," *IEEE Journal on Robotics and Automation*, vol. 3, pp. 239–248, June 1987.
- [14] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondance," in *Proceedings of the European Conference of Computer Vision 94*, pp. 151–158, 1994.