Stereo Correspondence from Motion Correspondence

F. Dornaika and R. Chung
Department of Mechanical and Automation Engineering
The Chinese University of Hong Kong, NT, Hong Kong
{dornaika, rchung}@mae.cuhk.edu.hk

Abstract

This paper introduces a new framework for stereo correspondence recovery using one motion of a stereo rig. Both the stereo correspondence and the motion of the stereo rig are unknown. By combining the stereo geometry and the motion correspondence we are able to infer the stereo correspondence from motion correspondence without having to systematically use the intensity-based stereo matching algorithms. The stereo correspondence recovery consists of two consecutive steps: the first step uses metric data associated with the stereo rig while the second step uses feature correspondences only. Experiments involving real stereo pairs indicate the feasibility and robustness of the approach.

1 Introduction

One of the most interesting goals of computer vision is the 3D structure recovery of scenes. This recovery has many applications such as object recognition and modeling, automatic cartography, and autonomous robot navigation. Traditionally, two cues have been used to obtain the 3D structure: i) *structure from motion* [6] and ii) *structure from stereo* [4], [1]. Both cues requires solutions of two subproblems: i) the correspondence problem and ii) the reconstruction problem. The correspondence problem (in stereo or motion) is one of the bottlenecks in computer vision. The motion cue has the advantage that the correspondence problem is relatively easy to solve because consecutive images are "alike", but the retrieved 3D structure is usually inaccurate due to the fact that the base line between views is so small. Consequently, the estimated depth will be inaccurate in the presence of image noise.

It is well recognized that the stereo cue has the advantage that the separation of the two views is significant enough to obtain accurate 3D reconstruction, however it has a difficult correspondence problem especially when the two views have many features. Several factors make the stereo correspondence problem difficult: occlusions, large disparities, photometric and figural distortions, a significant orientation of the image planes, deciding what window size to use in intensity-based matching. After all, establishing stereo correspondences is difficult because of the large search space for correspondence even with the epipolar geometry available.

In this paper, we introduce a new approach for binocular matching constraints. These constraints considerably reduce the stereo correspondence ambiguity. We consider two stereo pairs of the same scene. As an arbitrary number of intermediate frames can be available, monocular correspondences are easy to obtain. More precisely, we will show how we can combine the motion correspondences and the stereo geometry to infer the stereo correspondences using geometrical constraints. The method can be summarized as follows. First, the stereo rig motion is completely recovered. Second, the stereo correspondence are inferred from motion correspondences using two consecutive steps: i) the first step uses the stereo rig motion to get some initial stereo matches, ii) the second step uses these initial stereo matches to recover additional stereo matches. The advantage of the last step is that it does not depend on metric data (calibration and motion of the stereo rig) since it uses point transfer between images. Consequently, it will be more accurate to overcome the ambiguity of the stereo correspondence problem. Figure 1 displays an overview of the developed method.

This paper is organized as follows. Section 2 defines the problem we focus on. Section 3 describes the recovery of the stereo rig motion from motion correspondences in left and right images. Section 4 presents the recovery of some initial stereo correspondences. Section 5 presents a refinement of the stereo correspondence recovery based on feature correspondences only. Experimental results are presented in Section 6.

2 Problem statement

The problem is illustrated in Figure 2. A calibrated binocular stereo rig is considered. Each camera is described by a pin-hole model. This means that the pro-

1063-6919/99 $10.00 (c) 1999 IEEE

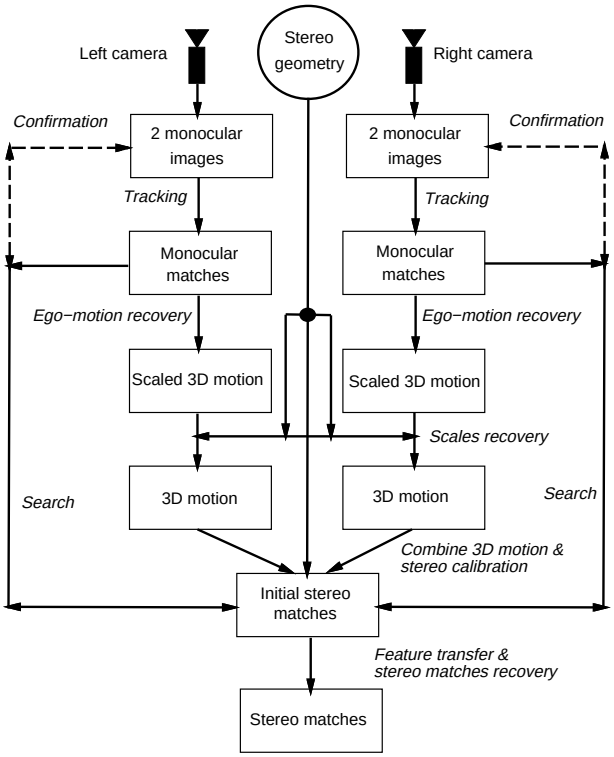


Figure 1: Illustration of the developed method.

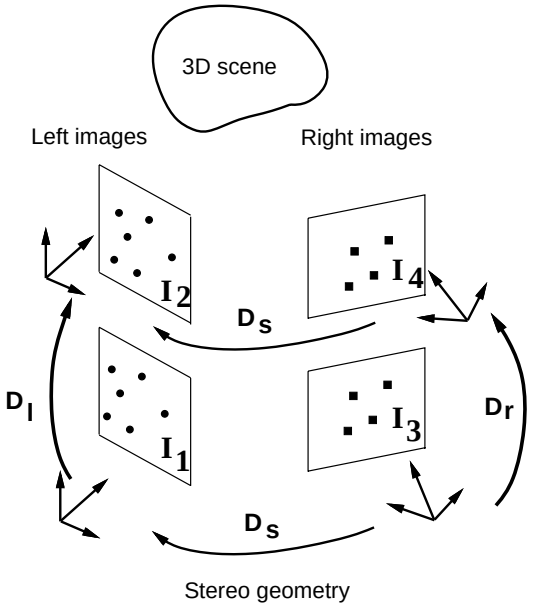


Figure 2: Illustration of the problem. Knowing the stereo geometry and motion correspondences in two stereo pairs, we like to constraint the stereo correspondences.

jection of the 3D scene onto 2D images will be described by a full perspective projection. In the sequel, image points as well as image lines will be represented by their homogeneous coordinates (a 3-vector). We assume that the stereo rig undergoes one single motion in front of an unknown scene. The obtained images are denoted by I_1 and I_2 for the left camera, and by I_3 and I_4 for the right camera. Image features are denoted by $\mathbf{m}^1$, $\mathbf{m}^2$, $\mathbf{m}^3$, and $\mathbf{m}^4$, in I_1, I_2, I_3, and I_4, respectively.

We denote by $\mathbf{K}_l$ ($\mathbf{K}_r$) the 3×3 calibration matrix of the left (right) camera. Both are upper triangular matrices. Let $\mathbf{D}_s$ be the Euclidean transformation of the stereo rig. $\mathbf{D}_s$ takes the form $\begin{bmatrix} \mathbf{R}_s & \mathbf{t}_s \\ \mathbf{0} & 1 \end{bmatrix}$ where $\mathbf{R}_s$ and $\mathbf{t}_s$ represent the rotation and the translation vector between the left camera and the right camera. Since the stereo rig is calibrated $\mathbf{K}_l$, $\mathbf{K}_r$, $\mathbf{R}_s$, and $\mathbf{t}_s$ are known *a priori*. Let $\mathbf{D}_l$ and $\mathbf{D}_r$ be the motion of the left and right camera (both are unknown and have the same form as $\mathbf{D}_s$). It is obvious that the stereo rig motion is fully described by either $\mathbf{D}_l$ or $\mathbf{D}_r$ since we have (see Figure 2):

$$\mathbf{D}_l \, \mathbf{D}_s = \mathbf{D}_s \, \mathbf{D}_r \tag{1}$$

By applying classic tracking methods to the monocular images, we can establish point-to-point correspondence between I_1 and I_2, and between I_3 and I_4, independently. The task of tracking can be made easier if we use intermediate frames between the two positions of the stereo rig. Hence, we obtain 2 different sets of feature points: the left ones and the right ones. We point out that the features in these 2 sets are not in one-to-one correspondence since they are obtained independently.

Our goal is to establish the matching between corresponding features in the 2 sets of features, i.e. stereo matching between images I_1 and I_3 (equivalently between I_2 and I_4).

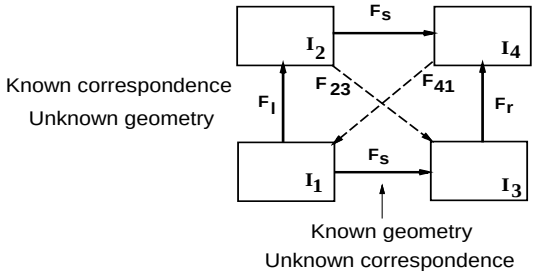


Figure 3: The stereo rig fundamental matrix $\mathbf{F}_s$ is recovered from the stereo geometry. The motion fundamental matrices $\mathbf{F}_l$ and $\mathbf{F}_r$ are recovered from point-to-point correspondences. $\mathbf{F}_{23}$ and $\mathbf{F}_{41}$ are unknown.

3 Computing the stereo rig motion

As depicted in Figure 3, it is obvious that on one hand the correspondence in monocular images is known but the 3D motion is not known. On the other hand, the correspondence in stereo images is not known but we know the 3D relation between the two cameras which is represented by $\mathbf{D}_s$. In this section, we will show that the 3D motion of the stereo rig can be completely recovered by combining the motion correspondences and the stereo geometry.

Let $\mathbf{F}_s$ be the fundamental matrix of the stereo rig, and $\mathbf{F}_l$ and $\mathbf{F}_r$ the fundamental matrices associated with the left and right motions. $\mathbf{F}_s$ is known from the stereo calibration. One can notice that fundamental matrices can be very efficiently and robustly estimated from point matches, given at least 8 point-to-point correspondences [3], [8]. Therefore, $\mathbf{F}_l$ and $\mathbf{F}_r$ can be recovered using the monocular correspondences ($I_1 \leftrightarrow I_2$) and ($I_3 \leftrightarrow I_4$), respectively (see Figure 3). In the remainder of this section we will show how the camera motion $\mathbf{D}_l$ and $\mathbf{D}_r$ can be completely recovered by combining the stereo geometry and the motion fundamental matrices.

3.1 Computing the scaled motions

It is recognized that if we know the intrinsic camera parameters then the camera motion (the rotation and the translation up to a scale factor) can be derived from point-to-point correspondences, i.e. from the corresponding fundamental matrix. This problem has been thoroughly studied in the literature [7], [9]. Closed-form solutions as well as non-linear solutions can be used. Therefore, the left motion $\mathbf{D}_l$ (the right motion $\mathbf{D}_r$) (up to a scale factor) can be recovered from the fundamental matrix $\mathbf{F}_l$ ($\mathbf{F}_r$).

3.2 Computing the scale factors

At this stage, we know the left (right) motion $\mathbf{D}_l$ ($\mathbf{D}_r$) up to a scale factor. Let λ_l and λ_r be the norm of the translation vectors $\mathbf{t}_l$ and $\mathbf{t}_r$, respectively. We have $\mathbf{t}_l = \lambda_l \hat{\mathbf{t}}_l$ and $\mathbf{t}_r = \lambda_r \hat{\mathbf{t}}_r$. The unit vectors $\hat{\mathbf{t}}_l$ and $\hat{\mathbf{t}}_r$ are known (Section 3.1). Equation (1) can be written as:

$$\begin{bmatrix} \mathbf{R}_l & \lambda_l \hat{\mathbf{t}}_l \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{bmatrix} \mathbf{R}_s & \mathbf{t}_s \\ \mathbf{0}^T & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R}_s & \mathbf{t}_s \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{bmatrix} \mathbf{R}_r & \lambda_r \hat{\mathbf{t}}_r \\ \mathbf{0}^T & 1 \end{bmatrix}$$

The translational part of this matrix equation is given by:

$$\mathbf{R}_l \, \mathbf{t}_s + \lambda_l \hat{\mathbf{t}}_l = \mathbf{R}_s \, \lambda_r \hat{\mathbf{t}}_r + \mathbf{t}_s \qquad (2)$$

Therefore, we have 3 linear equations in λ_l and λ_r. By solving this linear system we can recover λ_l and λ_r provided that $\mathbf{R}_l \neq \mathbf{I}$. $\mathbf{I}$ represents the 3×3 identity matrix. Consequently, the two motions $\mathbf{D}_l$ and $\mathbf{D}_r$ are completely recovered. Thus, it is possible to do complete 3D Euclidean reconstruction (scale included) of monocular features without having to solve for stereo correspondence.

4 Getting some initial stereo matches

Once the stereo rig motion is recovered by combining the motion correspondence and the stereo geometry, the fundamental matrices $\mathbf{F}_{23}$ and $\mathbf{F}_{41}$ will be straightforward ($\mathbf{S}()$ represents the skew-symmetric matrix associated with a 3-vector):

$$\mathbf{F}_{23} \cong \mathbf{K}_r^{-T} \mathbf{S}(\mathbf{t}_{23}) \mathbf{R}_{23} \mathbf{K}_l^{-1}$$

$$\mathbf{F}_{41} \cong \mathbf{K}_l^{-T} \mathbf{S}(\mathbf{t}_{41}) \mathbf{R}_{41} \mathbf{K}_r^{-1}$$

where $\mathbf{R}_{23}$, $\mathbf{t}_{23}$, $\mathbf{R}_{41}$, and $\mathbf{t}_{41}$ are derived from $\mathbf{D}_l$ and $\mathbf{D}_s$. Each left correspondence $\mathbf{m}^1 \leftrightarrow \mathbf{m}^2$ will have its stereo match (in image I_3) at the intersection of these two epipolar lines: $\mathbf{F}_s \, \mathbf{m}^1$ and $\mathbf{F}_{23} \, \mathbf{m}^2$. Therefore, a predicted location can be estimated by $(\mathbf{F}_s \, \mathbf{m}^1) \times (\mathbf{F}_{23} \, \mathbf{m}^2)$.

Thus, by searching in a small neighborhood among the extracted features in image I_3 one can detect the stereo match of $\mathbf{m}^1$. Thus, we are left with a set of stereo correspondences $\{\mathbf{s}^1 \leftrightarrow \mathbf{s}^3\}$. In order to eliminate possible mismatches within this set, a normalized correlation can be used. If we denote two $W \times W$ areas centered on features $\mathbf{s}^1$ and $\mathbf{s}^3$ as two arrays of pixel intensities $\mathbf{I}_1(uv)$ and $\mathbf{I}_3(uv)$, respectively, the normalized correlation is defined as:

$$C_{\mathbf{s}^1,\mathbf{s}^3} = \frac{\sum_{u=1}^{W} \sum_{v=1}^{W} (\mathbf{I}_1(uv) - \overline{\mathbf{I}}_1)(\mathbf{I}_3(uv) - \overline{\mathbf{I}}_3)}{W^2 \, \sigma(\mathbf{I}_1) \, \sigma(\mathbf{I}_3)}$$

where $\overline{\mathbf{I}}_1$ ($\overline{\mathbf{I}}_3$) is the average and $\sigma(\mathbf{I}_1)$ ($\sigma(\mathbf{I}_3)$) the standard deviation of all elements of the two areas. This coefficient varies from -1 for completely uncorrelated patches to 1 for identical patches.

Eventually, we have a set of stereo correspondences between the images I_1 and I_3 (equivalently between I_2 and I_4).

Alternatively, the initial stereo matches search can be also guided by using a simple 3D Euclidean reconstruction obtained from the two images I_1 and I_2, followed by a projection into the image I_3 since the relative displacements $\mathbf{D}_l$ and $\mathbf{D}_s$ are known. This is very useful in some cases where the concatenation of the epipolar geometry fails. In the next section, we will show that using these initial stereo matches, additional stereo correspondences can be inferred from motion correspondences without the use of the metric data associated with the stereo rig (calibration and 3D motion). The key idea relies on transferring the motion correspondences of one view to the images of the other view. Two equivalent methods can be used: the projective structure-based method and the trifocal tensor-based method.

5 Stereo correspondences recovery

5.1 Projective structure

The projective structure is nothing but the Euclidean structure distorted by a 4×4 projective transformation [2], [5]. It has the advantage that it can be computed using point-to-point correspondences only. To obtain a 3D projective structure of monocular features, we used the fundamental matrix based-method. Therefore, using motion correspondences in images I_1 and I_2 we can infer their 3D projective structure from the fundamental matrix $\mathbf{F}_l$. Similarly, using motion correspondences in images I_3 and I_4 we can infer their 3D projective structure from the fundamental matrix $\mathbf{F}_r$.

5.2 Projective structure–to–image plane mapping

The goal of this paragraph is to show how we can relate the left projective structure computed from I_1 and I_2 to the image plane I_3. This is possible through the recovery of a 3×4 mapping between the projective structure and the image plane I_3.

We denote by $\mathbf{s}_i^1 \leftrightarrow \mathbf{s}_i^2 \leftrightarrow \mathbf{s}_i^3 \leftrightarrow \mathbf{s}_i^4, i = 1 \cdots m$ the initial stereo correspondences obtained in Section 4. Let $\mathbf{S}_i$ (4-vector) be the projective coordinates of the corresponding 3D points (left ones). It is straightforward that we have m 3D to 2D point-to-point correspondences between the left projective structure and the right image I_3.

Let $\mathbf{P}_l$ be the 3×4 projective matrix describing the mapping between the left projective structure and I_3 (see Figure 4). Thus we have the following constraints ($i = 1 \cdots m$):

$$\mathbf{s}_i^3 \cong \mathbf{P}_l \, \mathbf{S}_i \tag{3}$$

Since each point correspondence provides two linear equations in the entries of the mapping $\mathbf{P}_l$, we can conclude that $\mathbf{P}_l$ can be linearly estimated from (3) provided that the number of initial stereo correspondences is equal to or greater than 6 ($m \geq 6$). Similarly, one can estimate the mapping $\mathbf{P}_r$ between the right projective structure and the left image I_1.

5.3 Stereo correspondences recovery

We consider any motion correspondence $\mathbf{m}^1 \leftrightarrow \mathbf{m}^2$ in the left images I_1 and I_2. Let $\mathbf{M}$ be the 3D projective coordinates of this point (Section 5.1). The 2D location $\mathbf{m}^{3\star}$ of its stereo correspondence in the right image I_3 can be predicted by (see Figure 4):

$$\mathbf{m}^{3\star} \cong \mathbf{P}_l \, \mathbf{M} \tag{4}$$

In the presence of image noise, $\mathbf{m}^{3\star}$ will not coincide with the real location of the stereo match of $\mathbf{m}^1$. Therefore, we use this location together with the epipolar line

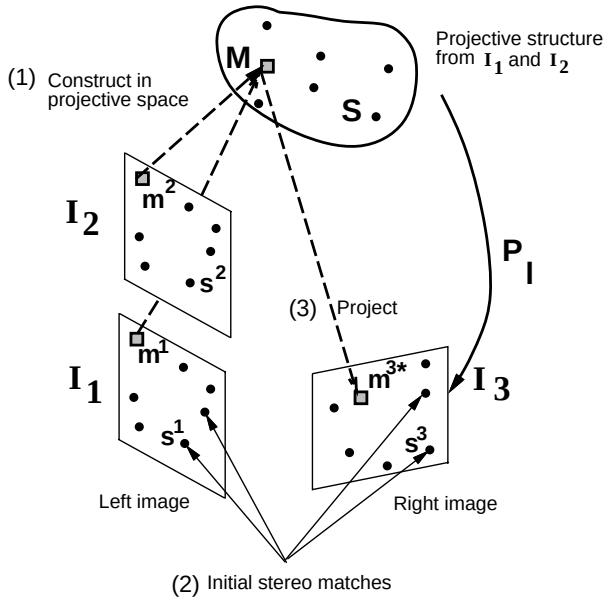


Figure 4: Prediction of the 2D location of the corresponding feature. First, the left feature is projectively reconstructed (1). Second, it is projected into the right image using the mapping between the left projective structure and the right image (3) (this mapping is recovered from a few initial stereo correspondences (2)).

(represented by $\mathbf{F}_s \, \mathbf{m}^1$) to determine a small neighborhood in I_3 in which one expects to find the stereo match of $\mathbf{m}^1$ (see Figure 5). Since feature extraction is done initially in a separate manner for the left and right images, some of the tracked features in I_1 and I_2 may not have the corresponding match in I_3. Therefore, if some right extracted features belong to the defined neighborhood, the stereo match of $\mathbf{m}^1$ will be the closest feature to the associated epipolar line.

Similarly, one can infer stereo correspondences using the right projective structure obtained from I_3 and I_4. In practice, since we put extracted features into correspondence using geometrical constraint, a one way stereo correspondence is sufficient, i.e. either left $\longrightarrow$ right or right $\longrightarrow$ left.

6 Experiments

The proposed method was applied to real stereo pairs. The first experiment is performed on 2 stereo pairs of a natural scene in which many rocks are gathered on a rectangular platform (see Figure 6). The 2 cameras have a baseline of 38 cm. 250 feature points are tracked in the left two images, and 288 feature points are tracked in the right images. The entire motion of the stereo rig is about 37 cm. Using these features the left and right fundamental matrices are

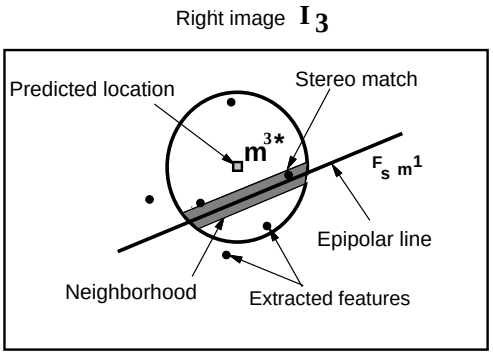


Figure 5: The predicted location $\mathbf{m}^{3\star}$ together with the epipolar line define a small neighborhood in the right image I_3. Therefore, the stereo match will be the extracted feature which belongs to this neighborhood. In case we have more than one, the stereo match will be the closest feature to the epipolar line.

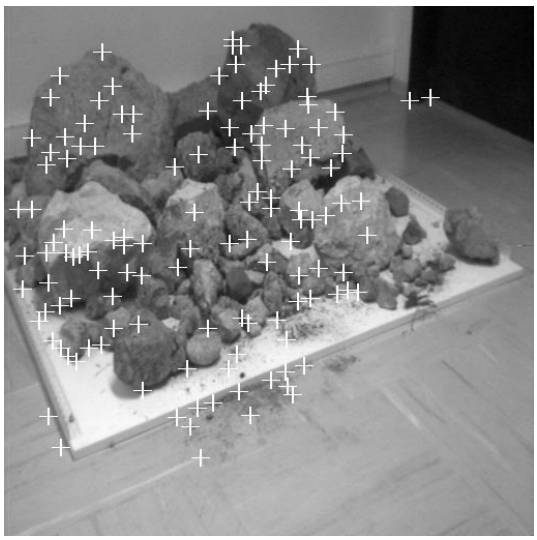
computed. Then, the stereo rig motion is recovered as explained in Section 3. Using this motion a total of 16 initial stereo matches were located in the left and right images (Section 4) . By applying the projective structure-based method (transfer method), we can find 132 successful stereo matches (Section 5). Figure 6 shows the recovered stereo matches (the first 2 images) while the bottom displays a top view of the reconstructed stereo matches. It takes 0.45 second on an Ultra-Sparc to carry out the stereo correspondence recovery as it is described in Section 5. In a second experiment we have two stereo pairs of a house model (see Figure 7). The stereo rig motion is about 12 cm. A total of 13 initial stereo matches were located in the left and right images. By applying the projective structure-based method, we can find 176 successful stereo matches whose positions are shown in Figure 7. This figure (bottom) illustrates a top view of the reconstructed features. One can notice that the two faces of the house as well as the two chimneys are reconstructed correctly.

It should be noticed that in both experiments, the stereo matches are obtained from extracted features (motion correspondences) using only geometric constraints. One can expect to get a few extra stereo matches if the constraints will be combined with any stereo correlation-based search.

7 Conclusion

In this paper, we have investigated the stereo correspondence problem using one motion of a stereo rig. We have developed binocular matching constraints that takes advantage of the motion correspondences and the knowledge of the stereo geometry. It has been argued that the mo-

left image (I_1)



right image (I_3)

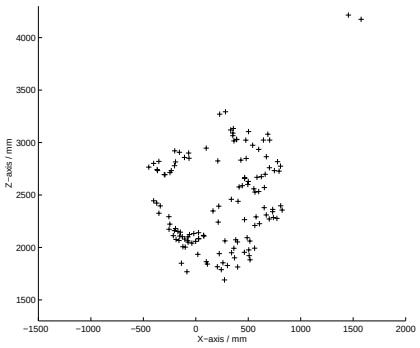


Figure 6: The first stereo pairs and the recovered stereo matches (the top 2 images). Top view of the reconstructed stereo matches (Bottom).

left image (I_1)

right image (I_3)

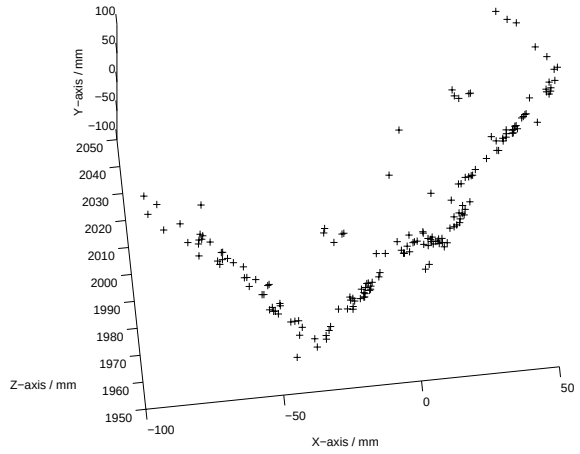


Figure 7: The first stereo pairs and the recovered stereo matches (the top 2 images). Top view of the reconstructed stereo matches (Bottom).

tion correspondence are easier to obtain. First, the stereo rig motion is recovered by combining the stereo geometry and the motion correspondences. Second, the stereo correspondences are inferred from motion correspondences using two consecutive steps: i) the first step uses the stereo rig motion to infer some stereo correspondences, and ii) the second step uses these ones to infer additional stereo correspondences through a point transfer. Certainly the camera motion cannot be too close to zero. However, tracking of monocular features can be aided by the use of many intermediate frames. The developed method can be useful to many stereo matching algorithms. Moreover, it can be used to completely recover the 3D shape (not just the scaled one) using monocular matches only.

Acknowledgment

The work described in this paper was substantially supported by a grant from the Research Grants Council of Hong Kong Special Administrative Region, China (RGC Ref. No. CUHK4114/97/E). It was also partially supported by the CUHK Postdoctoral Fellowship Scheme 1998 (CUHK Ref. No. 98/15/ERG).

References

[1] U. R. Dhond and J. K. Aggarwal. Structure from stereo - A review. *IEEE Trans. System., Man. & Cybern.*, 19(6):1489–1510, 1989.

[2] O. Faugeras. *"Three-Dimensional Computer Vision: a Geometric Viewpoint"*. The MIT Press, 1993.

[3] R. I. Hartley. In defence of the 8-point algorithm. In *Proc. of the International Conference on Computer Vision*, pages 1064–1070, June 1995.

[4] N. Navab, Z. Zhang, and O. Faugeras. Tracking, motion and stereo: A robust and dynamic cooperation. In *Proc. Scandinavian Conference on Image Analysis*, pages 98–105, August 1991.

[5] C. Rothwell, G. Csurka, and O. Faugeras. A comparison of projective reconstruction methods for pairs of views. In *Proc. of the International Conference on Computer Vision*, pages 932–937, 1995.

[6] K. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[7] J. Weng, T. S. Huang, and N. Ahuja. *Motion and Structure from Image Sequences*. Springer-Verlag, Berlin, 1993.

[8] Z. Zhang. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision*, 27(2):43–76, 1998.

[9] Z. Zhang. On the optimization criteria used in two-view motion analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(7):717–729, 1998.