# Accurate Scale Factor Estimation in 3D Reconstruction*

Manolis Lourakis and Xenophon Zabulis

Institute of Computer Science,
Foundation for Research and Technology - Hellas (FORTH)
Vassilika Vouton, P.O.Box 1385, GR 711 10, Heraklion, Crete, Greece

**Abstract.** A well-known ambiguity in monocular structure from motion estimation is that 3D reconstruction is possible up to a similarity transformation, i.e. an isometry composed with isotropic scaling. To raise this ambiguity, it is commonly suggested to manually measure an absolute distance in the environment and then use it to scale a reconstruction accordingly. In practice, however, it is often the case that such a measurement cannot be performed with sufficient accuracy, compromising certain uses of a 3D reconstruction that require the acquisition of true Euclidean measurements. This paper studies three alternative techniques for obtaining estimates of the scale pertaining to a reconstruction and compares them experimentally with the aid of real and synthetic data.

**Keywords:** structure from motion, scale ambiguity, pose estimation.

## 1 Introduction

Structure from motion with a single camera aims at recovering both the 3D structure of the world and the motion of the camera used to photograph it. Without any external knowledge, this process is subject to the inherent scale ambiguity [9,17,5], which consists in the fact that the recovered 3D structure and the translational component of camera motion are defined up to an unknown scale factor which cannot be determined from images alone. This is because if a scene and a camera are scaled together, this change would not be discernible in the captured images. However, in applications such as robotic manipulation or augmented reality which need to interact with the environment using Euclidean measurements, the scale of a reconstruction has to be known quite accurately.

Albeit important, scale estimation is an often overlooked step by structure from motion algorithms. It is commonly suggested that scale should be estimated by manually measuring a single absolute distance in the scene and then using it to scale a reconstruction to its physical dimensions [5,12]. In practice, there are two problems associated with such an approach. The first is that it favors certain elements of the reconstruction, possibly biasing the estimated scale. The second, and more important, is that the distance in question has to be measured

* Work funded by the EC FP7 programme under grant no. 270138 DARWIN.

R. Wilson et al. (Eds.): CAIP 2013, Part I, LNCS 8047, pp. 498–506, 2013.

accurately in the world and then correctly associated with the corresponding distance in the 3D reconstruction. Such a task can be quite difficult to perform and is better suited to large-scale reconstructions for which the measurement error can be negligible compared to the distance being measured. However, measuring distances for objects at the centimeter scale has to be performed with extreme care and is therefore remarkably challenging. For example, [1] observes that a modeling error of $1mm$ in the scale of a coke can, gives rise to a depth estimation error of up to $3cm$ at a distance of $1m$ from the camera, which is large enough to cause problems to a robotic manipulator attempting to grasp the object.

This work investigates three techniques for obtaining reliable scale estimates pertaining to a monocular 3D reconstruction and evaluates them experimentally. These techniques differ in their required level of manual intervention, their flexibility and accuracy. Section 2 briefly presents our approach for obtaining a reconstruction whose scale is to be estimated. Scale estimation techniques are detailed in Sections 3-5 and experimental results from their application to real and synthetic datasets are reported in Sect. 6. The paper concludes in Sect. 7.

## 2 Obtaining a 3D Reconstruction

In this work, 3D reconstruction refers to the recovery of sparse sets of points from an object's surface. To obtain a complete and view independent representation, several images depicting an object from multiple unknown viewpoints are acquired with a single camera. These images are used in a feature-based structure from motion pipeline to estimate the interimage camera motion and recover a corresponding 3D point cloud [16]. This pipeline relies on the detection and matching of SIFT keypoints across images which are then reconstructed in 3D. The 3D coordinates are complemented by associating with each reconstructed point a SIFT feature descriptor [11], which captures the local surface appearance in the point's vicinity. A SIFT descriptor is available from each image where a particular 3D point is seen. Thus, we select as its most representative descriptor the one originating from the image in which the imaged surface is most frontal and close enough to the camera. This requires knowledge of the surface normal, which is obtained by gathering the point's 3D neighbours and robustly fitting to them a plane. As will become clear in the following, SIFT descriptors permit the establishment of putative correspondences between an image and an object's 3D geometry. Combined together, 3D points and SIFT descriptors of their image projections constitute an object's representation.

## 3 Scale Estimation from Known Object Motion

The simplest approach to estimate an object's scale employs a single static camera to acquire two views of the object in different poses with known relative displacement. Then, the pose of the object in each view is determined. Since the camera is static, the two poses estimated can be used to compute the object's displacement up to the unknown scale. The sought scale is simply the ratio of known over recovered displacement. To ease the task of measuring 3D displacements, the object is placed so that it is aligned with the checkers of a checkerboard grid. Such a guided placemement allows the distance between the object's locations to be known through the actual size of each checker. An advantage of the known motion approach is that it does not involve a special camera setup. On the other hand, it suffers from two disadvantages. First, it relies on careful object placement on the grid and is, therefore, susceptible to human error. Second, it treats images separately and thus does not avail any opportunities for combining them and in so doing increase the overall accuracy.

A key ingredient of the method outlined above is the estimation of the pose of a known object in a single image, therefore more details regarding this computation are provided next. Given an image of the object, SIFT keypoints are detected in it and then matched against those contained in its reconstructed representation (cf. Sect. 2). The invariance of SIFT permits the reliable identification of features that have undergone large affine distortions in the image. The established correspondences are used to associate the 2D image locations of detected features with the 3D coordinates of their corresponding points on the object's surface. The procedure adopted for point matching is the F2P strategy from [8]. Compared to the standard test defined by the ratio of the distances to the closest and second closest neighbors [11], F2P was found to yield fewer erroneous matches. An important detail concerns the quantification of distances among SIFT descriptors, which are traditionally computed with the Euclidean ($L_2$) norm. Considering that the SIFT descriptor is a weighted histogram of gradient orientations, improvements in matching are attained by substituting $L_2$ with histogram norms such as the Chi-squared ($\chi^2$) distance [15]. This is a histogram distance that takes into account the fact that in many natural histograms, the difference between large bins is less important than the difference between small bins and should therefore be reduced. Keypoint matching provides a set of 3D-2D correspondences from which pose is estimated as explained below.

Pose estimation concerns determining the position and orientation of an object with respect to a camera given the camera intrinsics and a set of $n$ correspondences between known 3D object points and their image projections. This problem, also known as the Perspective-n-Point (PnP) problem, is typically solved using non-iterative approaches that involve small, fixed-size sets of correspondences. For example, the basic case for triplets ($n = 3$, known as the P3P problem), has been studied in [3] whereas other solutions were later proposed in [2,7]. P3P is known to admit up to four different solutions, whereas in practice it usually has just two. Our approach for pose estimation in a single image uses a set of 2D-3D point correspondences to compute a preliminary pose estimate and then refine it iteratively. This is achieved by embedding the P3P solver [3] into a RANSAC [2] framework and computing an initial pose estimate along with a classification of correspondences into inliers and outliers. The pose computed by RANSAC is next refined to take into account all inlying correspondences by minimizing a non-linear cost function corresponding to their total reprojection error. The minimization is made more immune to noise caused by mislocalized image points by substituting the squared reprojection error with a robust cost function (i.e., M-estimator). Our pose estimation approach is detailed in [10].

## 4 Scale Estimation from 3D Reconstruction and Absolute Orientation

Another way of approaching the scale estimation problem is to resort to stereo. More specifically, a strongly calibrated stereo pair is assumed and two-view triangulation is employed to estimate the 3D coordinates of points on the surface of the object. These points are then matched to points from the object's representation. The scale factor is estimated by finding the similarity aligning the triangulated 3D points with their counterparts from the representation. This is achieved by solving the absolute orientation problem, which also accounts for the unknown scale. To safeguard against possible outliers, the calculation is embedded in a RANSAC robust estimation scheme that seeks the transformation aligning together a fraction of the available 3D matches. More details regarding the solution of the absolute orientation problem are given next.

Starting with a stereo image pair depicting the object whose scale is to be estimated, sparse correspondences between the two images are established. This is achieved by detecting SIFT features in each image and then matching them through their descriptors. For each pair of corresponding points, stereo triangulation is used to estimate the 3D coordinates of the imaged world point [4]. Knowledge of the extrinsic calibration of the stereo rig permits the triangulated points to be expressed in their true scale. Further to their matching in the stereo images, SIFT descriptors are also matched against the descriptors stored in the representation. In other words, three-way correspondences are established between object points in the two images and the representation. In this manner, the triangulated points are associated with 3D points from the object's representation. The sought scale factor is then computed by determining the similarity between the triangulated 3D points and their counterparts, as follows.

Let $\{\mathbf{M}_i\}$ be a set of $n \geq 3$ reference points from the representation expressed in an object-centered reference frame and $\{\mathbf{N}_i\}$ a set of corresponding camera-space triangulated points. Assume also that the two sets of points are related by a similarity transformation as $\mathbf{N}_i = \lambda \, \mathbf{R} \, \mathbf{M}_i \, + \, \mathbf{t}$, where $\lambda$ is the sought scale factor and $\mathbf{R}$, $\mathbf{t}$ a rotation matrix and translation vector defining an isometry. As shown by Horn [6], absolute orientation can be solved using at least three non-collinear reference points and singular value decomposition (SVD). The solution proceeds by defining the centroids $\overline{\mathbf{M}}$ and $\overline{\mathbf{N}}$ and the locations $\{\mathbf{M}_i^{'}\}$ and $\{\mathbf{N}_i^{'}\}$ of 3D points relative to them:

$$\overline{\mathbf{M}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{M}_i \ , \ \overline{\mathbf{N}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{N}_i \ , \quad \mathbf{M}_i^{'} = \mathbf{M}_i - \overline{\mathbf{M}} \ , \ \mathbf{N}_i^{'} = \mathbf{N}_i - \overline{\mathbf{N}}.$$

Forming the cross-covariance matrix $\mathbf{C}$ as $\sum_{i=1}^{n} \mathbf{N}_i^{'} \, \mathbf{M}_i^{'t}$, the rotational component of the similarity is directly computed from $\mathbf{C}$'s decomposition $\mathbf{C} = \mathbf{U} \, \mathbf{\Sigma} \, \mathbf{V}^t$

as $\mathbf{R} = \mathbf{V}\ \mathbf{U}^t$. The scale factor is given by

$$\lambda = \sqrt{\sum_{i=1}^{n} ||\mathbf{M}'_i||^2 \Big/ \sum_{i=1}^{n} ||\mathbf{N}'_i||^2}\ , \tag{1}$$

whereas the translation follows as $\mathbf{t} = \overline{\mathbf{N}} - \lambda\, \mathbf{R}\, \overline{\mathbf{M}}$.

The primary advantages of this method over the one of Sect. 3 are that it does not require a particular object positioning strategy nor the measurement of any distances. Any object placement, provided that it is well imaged and avails sufficient correspondences, is suitable for applying the method. On the other hand, 3D reconstruction of points based on binocular stereo is often error-prone [13] and such inaccuracies can significantly affect the final estimation result.

## 5 Scale Estimation from Binocular Reprojection Error

Similarly to that in Sect. 4, this method also employs an extrinsically calibrated stereo pair. Given an object's 3D representation, its scale is determined by considering the reprojection error pertaining to the object's projections in the two images. Using the same coordinate system for both cameras, the reprojection error is expressed by an objective function which also includes scale in addition to rotation and translation. Then, the object's scale and pose are jointly estimated by minimizing the total reprojection error in both images, as follows.

The method starts by detecting SIFT keypoints in both stereo images. Independently for each image, the extracted keypoints are matched against the points of the representation through their descriptors. For each image, monocular pose estimation is carried out as described in Sect. 3 to determine the object's pose in it. Knowledge of the camera extrinsics allows us to express both of these poses in the same coordinate system, for example that of the left camera. Indeed, if the pose of the object in the left camera is defined by $\mathbf{R}$ and $\mathbf{t}$, its pose in the right camera equals $\mathbf{R_s R}$ and $\mathbf{R}_s\mathbf{t} + \mathbf{t}_s$, where $\mathbf{R_s}$ and $\mathbf{t_s}$ correspond to the pose of the right camera with respect to the left. Due to the stereo rig being rigid, $\mathbf{R_s}$ and $\mathbf{t_s}$ remain constant and can be estimated offline via extrinsic calibration. The most plausible scale and left camera pose are determined via the minimization of the cumulative reprojection error in both images. The binocular reprojection error consists of two additive terms, one for each image. More specifically, denoting the intrinsics for the left and right images by $\mathbf{K}^L$ and $\mathbf{K}^R$, the binocular reprojection error for $n$ points in the left image and $m$ in the right is defined as:

$$\sum_{i=1}^{n} d(\mathbf{K}^L \cdot [\lambda\, \mathbf{R}(\mathbf{r}) \,|\, \mathbf{t}] \cdot \mathbf{M}_i - \mathbf{m}_i^L)^2 + \sum_{j=1}^{m} d(\mathbf{K}^R \cdot [\lambda\, \mathbf{R}_s \mathbf{R}(\mathbf{r}) \,|\, \mathbf{R}_s\mathbf{t} + \mathbf{t}_s] \cdot \mathbf{M}_j - \mathbf{m}_j^R)^2, \tag{2}$$

where $\lambda$, $\mathbf{t}$ and $\mathbf{R}(\mathbf{r})$ are respectively the sought scale factor, translation vector and rotation matrix parameterized using the Rodrigues rotation vector $\mathbf{r}$, $\mathbf{K}^L \cdot$

$[\lambda\ \mathbf{R}(\mathbf{r})\,|\,\mathbf{t}] \cdot \mathbf{M}_i$ is the projection of homogeneous point $\mathbf{M}_i$ in the left image, $\mathbf{K}^R \cdot [\lambda\ \mathbf{R}_s\mathbf{R}(\mathbf{r})\,|\,\mathbf{R}_s\mathbf{t}+\mathbf{t}_s] \cdot \mathbf{M}_j$ is the projection of homogeneous point $\mathbf{M}_j$ in the right image, $\mathbf{m}_i^L$ and $\mathbf{m}_j^R$ are respectively the 2D points corresponding to $\mathbf{M}_i$ and $\mathbf{M}_j$ in the left and right images and $d(\mathbf{x}, \mathbf{y})$ denotes the reprojection error, i.e. the Euclidean distance between the image points represented by vectors $\mathbf{x}$ and $\mathbf{y}$. The expression in (2) can be extended to an arbitrary number of cameras and is minimized with respect to $\lambda$, $\mathbf{r}$, $\mathbf{t}$ with the Levenberg-Marquardt non-linear least squares algorithm, employing only the inliers of the two monocular estimations to ensure resilience to outliers. Similarly to the monocular case, a M-estimate of the reprojection error is minimized rather than the squared Euclidean norm. One possible initialization is to start the minimization from the monocular pose computed for the left camera. Still, this initialization does not treat images symmetrically as it gives more importance to the left image. Therefore, if the pose with respect to the left camera has been computed with less precision than that in the right, there is a risk of the binocular refinement also converging to a suboptimal solution. To remedy this, the refinement scheme is extended by also using the right image as reference and refining pose in it using both cameras, assuming a constant transformation from the left to the right camera. Then, the pose yielding the smaller overall binocular reprojection error is selected.

This method has several attractive features: It does not require a particular object placement strategy. There is no need for a short baseline as correspondences are not established across the two views but, rather, between each individual view and the reconstruction. Because no attempt is made to reconstruct in 3D, the experimental setup is relieved from the constraints related to the binocular matching of points and the inaccuracies associated with their reconstruction. A direct consequence of this is that the two cameras may have very different viewpoints. In fact, employing large baselines favours the method as it better constrains the problem of scale factor estimation.

## 6 Experiments

Each of the three methods previously described provides a means for computing a single estimate of the pursued scale factor through monocular or binocular measurements. It is reasonable to expect that such estimates will be affected by various errors, therefore basing scale estimation on a single pair of images should be avoided. Instead, more accurate estimates can be obtained by employing multiple images in which the object has been moved to different positions and collecting the corresponding estimates. Then, the final scale estimate is obtained by applying a robust location estimator such as their sample median [14]. In the following, the methods of Sect. 3, 4 and 5 will be denoted as MONO, ABSOR and REPROJ, respectively. Due to limited space, two sets of experiments are reported.

An experiment with synthetic images was conducted first, in which the baseline of the stereo pair imaging the target object was varied. A set of images was generated, utilizing a custom OpenGL renderer. A 1:1 model of a textured rectangular cuboid (sized $45 \times 45 \times 90\,mm^3$), represented by a 3D triangle mesh (with 14433 vertices & 28687 faces), was rendered in 59 images. These images correspond to a virtual camera (1280 × 960 pixels, 22.2° × 16.7° FOV) circumventing the object in a full circle of radius $500\,mm$ perpendicular to its major symmetry axis. At all simulated camera locations, the optical axis was oriented so that it pointed towards the object's centroid. The experiment was conducted in 30 conditions, each employing an increasingly larger baseline. In condition $n$, the $i^{th}$ stereo pair comprised of images $i$ and $i+n$. Hence, the baseline increment in each successive condition was $\approx 52mm$. In Fig. 1(a) and (b), an image from the experiments and the absolute error in the estimated scale factor are shown. Notice that the plot for ABSOR terminates early at a baseline of $\approx 209mm$. This is because as the baseline length increases, the reduction in overlap between the two images of the stereo pair results in fewer correspondences. In conditions of the experiment corresponding to larger baselines, some stereo pairs did not provide enough correspondences to support a reliable estimate by ABSOR. As a result, the estimation error for these pairs was overly large.

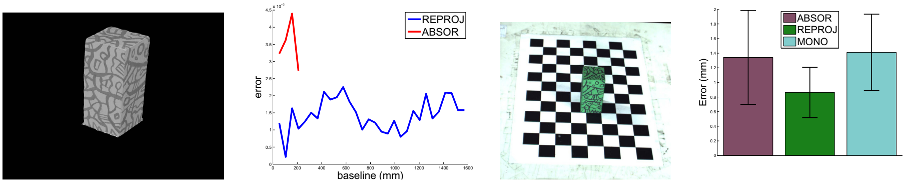

**Fig. 1.** Experiments. Left to right: (a) sample image from the experiment with synthetic stereo images and (b) scale factor estimation error (in milli scale), (c) sample image from the experiment with real images and (d) translational pose estimation error.

The three methods are compared next with the aid of real images. Considering that the task of directly using the estimated scales to assess their accuracy is cumbersome, it was chosen to compare scales indirectly through pose estimation. More specifically, an arbitrarily scaled model of an object was re-scaled with the estimates provided by MONO, ABSOR and REPROJ. Following this, these re-scaled models were used for estimating poses of the object as explained in Sect. 3, which were then compared with the true poses. In this manner, the accuracy of a scale estimate is reflected on the accuracy of the translational components of the estimated poses. To obtain ground truth for object poses, a checkerboard was used to guide the placement of the object that was systematically moved at locations aligned with the checkers. The camera pose with respect to the checkerboard was estimated through conventional extrinsic calibration, from which the locations of the object on the checkerboard were transformed to the camera reference frame. The object and the experimental setup are shown in Fig. 1(c). Note that these presumed locations include minute calibration inaccuracies as well as human errors in object placement. The object was placed and aligned upon every checker of the 8 × 12 checkerboard in the image. The checkerboard was at a distance of approximately $1.5\,m$ from the camera, with each checker being $32 \times 32\,mm^2$.

Camera resolution was $1280 \times 960$ pixels, and its FOV was $16° \times 21°$. The mean translational error in these 96 trials was $1.411\,mm$ with a deviation of $0.522\,mm$ for MONO, $1.342\,mm$ with a deviation of $0.643\,mm$ for ABSOR and $0.863\,mm$ with a deviation of $0.344\,mm$ for REPROJ. The mean translational errors of the pose estimates are shown graphically in Fig. 1(d).

## 7 Conclusion

The paper has presented one monocular and two binocular methods for scale factor estimation. Binocular methods are preferable due to their flexibility with respect to object placement. Furthermore, the binocular method of Sect. 5 is applicable regardless of the size of the baseline and was shown to be the most accurate, hence it constitutes our recommended means for scale estimation.

## References

1. Collet Romea, A., Srinivasa, S.: Efficient Multi-View Object Recognition and Full Pose Estimation. In: Proc. of ICRA 2010 (May 2010)
2. Fischler, M., Bolles, R.: RanSaC: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. In: CACM, vol. 24, pp. 381–395 (1981)
3. Grunert, J.: Das pothenotische Problem in erweiterter Gestalt nebst über seine Anwendungen in Geodäsie. Grunerts Archiv für Mathematik und Physik (1841)
4. Hartley, R., Sturm, P.: Triangulation. CVIU 68(2), 146–157 (1997)
5. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press (2004) ISBN: 0521540518
6. Horn, B.: Closed-form Solution of Absolute Orientation Using Unit Quaternions. J. Optical Soc. Am. A 4(4), 629–642 (1987)
7. Kneip, L., Scaramuzza, D., Siegwart, R.: A Novel Parametrization of the Perspective-three-Point Problem for a Direct Computation of Absolute Camera Position and Orientation. In: Proc. of CVPR 2011, pp. 2969–2976 (2011)
8. Li, Y., Snavely, N., Huttenlocher, D.P.: Location Recognition Using Prioritized Feature Matching. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 791–804. Springer, Heidelberg (2010)
9. Longuet-Higgins, H.: A Computer Algorithm for Reconstructing a Scene From Two Projections. Nature 293(5828), 133–135 (1981)
10. Lourakis, M., Zabulis, X.: Model-Based Pose Estimation for Rigid Objects. In: Chen, M., Leibe, B., Neumann, B. (eds.) ICVS 2013. LNCS, vol. 7963, pp. 83–92. Springer, Heidelberg (2013)
11. Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vis. 60(2), 91–110 (2004)
12. Moons, T., Gool, L.V., Vergauwen, M.: 3D Reconstruction from Multiple Images Part 1: Principles. Found. Trends. Comput. Graph. Vis. 4(4), 287–404 (2009)
13. Nistér, D., Naroditsky, O., Bergen, J.: Visual Odometry for Ground Vehicle Applications. J. Field Robot. 23, 3–20 (2006)
14. Rousseeuw, P.: Least Median of Squares Regression. J. Am. Stat. Assoc. 79, 871–880 (1984)

15. Rubner, Y., Puzicha, J., Tomasi, C., Buhmann, J.: Empirical Evaluation of Dissimilarity Measures for Color and Texture. Comput. Vis. Image Und. 84(1), 25–43 (2001)
16. Snavely, N., Seitz, S., Szeliski, R.: Photo Tourism: Exploring Photo Collections in 3D. ACM Trans. Graph. 25(3), 835–846 (2006)
17. Szeliski, R., Kang, S.: Shape Ambiguities in Structure from Motion. In: Buxton, B.F., Cipolla, R. (eds.) ECCV 1996. LNCS, vol. 1064, pp. 709–721. Springer, Heidelberg (1996)