Model-Based Pose Estimation for Rigid Objects*

Manolis Lourakis and Xenophon Zabulis

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

Abstract. Determining the pose of objects appearing in images is a problem encountered often in several practical applications. The most effective strategy for dealing with this challenge is to proceed according to the model-based paradigm, which involves building 3D models of objects and then determining object poses by fitting their models to new images with the aid of detected features. This paper proposes a model-based approach for estimating the full pose of known objects from natural point features. The method employs a projective imaging model and incorporates reliable automatic mechanisms for pose initialization and convergence. Furthermore, it is extendable to multiple cameras without the need to perform multi-view matching and relies on sparse structure from motion techniques for the construction of object models offline. Experimental results demonstrate its accuracy and robustness.

Keywords: Pose estimation, feature matching, object detection & recognition.

1 Introduction

Accurate localization of objects in images is a primary requirement for vision systems applied to areas such as robotic manipulation, tracking, augmented reality, tangible interfaces, etc. Such systems are expected to operate reliably in dynamic and unknown environments, delivering accurate object position and orientation estimates despite any variations in the appearance of objects due to changes in viewing position, illumination or occlusion. Object pose estimation is often addressed in the context of model-based matching and recognition [11], for which a vast body of literature is available. According to this paradigm, a collection of geometric object models and their associated features is assembled first. During recognition, features extracted from an image are matched to those stored in a model and the mapping among them is used to determine the pose of the corresponding object.

This paper presents a system for estimating the full, i.e. six degrees of freedom, pose of rigid objects. The system employs an offline stage to build a library of models encoding the 3D geometry and local appearance of objects, followed by an online stage for matching and pose estimation. Stored models consist of

* This work has received funding from the EC FP7 programme under grant no. 270138 DARWIN.

M. Chen, B. Leibe, and B. Neumann (Eds.): ICVS 2013, LNCS 7963, pp. 83–92, 2013.

sparse sets of 3D points from an object's surface along with the SIFT descriptors [13] of their image projections. During online operation, image features and their SIFT descriptors are extracted from images and matched against those of the models to establish putative correspondences. Owing to the specificity of SIFT, the number of these correspondences usually provides strong evidence regarding the presence of particular modelled objects in an image. Such hypotheses are tested by using the 3D coordinates of matched model points to estimate object poses and thus verify that correspondences occur in a configuration consistent with geometry. The proposed approach is applicable regardless of the relative pose of the object with respect to the camera, is robust to occlusions and mismatches and can easily recover from failures as it maintains very little state information. It makes mild assumptions regarding objects, postulating that they are rigid and textured but arbitrarily complex. Furthermore, it employs a fully projective imaging formulation, mechanisms for reliable pose initialization and convergence, is extendable to multiple cameras without the need to perform multi-view matching and relies on automatic structure from motion (SfM) techniques for the construction of offline models. Related existing work is reviewed in Sect. 2, the components of the proposed method are detailed in Sections 3-5 and experimental results from real and synthetic datasets are presented in Sect. 6.

2 Related Work

Early approaches to model-based matching employed intensity edges as features. Lowe [11], for example, relied on perceptual organization to group features and reduce the size of the search space involved in matching, followed by top-down spatial correspondence aimed at aligning a model with an image and estimating its pose. This work was confined to using straight line segments and employed approximate parallelism as their grouping cue. Since this property is not preserved under perspective, the method is limited to images with affine geometry for which the perspective distortion is small. The approach was subsequently extended to handle objects with arbitrary curved surfaces and internal parameters representing articulations or surface deformations [12].

Later on, developments in covariant detectors and descriptors for image patches [14] were adopted to build local representations. Thus, Vacchetti et al. [18] combine geometric models with feature-based matching against a set of reference keyframes to track rigid objects in 3D. Rothganger et al. [15] capture the appearance of object surface patches using affine invariant local descriptors and their spatial relationships using multi-view geometric constraints. Matching enforcing photometric and geometric consistency achieves object recognition and pose estimation. The approach assumes an affine projection model and incurs high computational cost. Gordon and Lowe [3] describe a system based on SIFT features for recognizing learnt models in new images and solving for their pose. Intended for use in an augmented reality application, this system estimates the pose of the camera with respect to a set of mostly stationary objects in its environment rather than the other way round. As a result, it puts emphasis on the

reduction of jitter and drift and can handle only a single model at a time that corresponds to the scene being tracked. The work of Collet et al. [1], who present a system based on natural features capable of estimating the pose of objects in a robot's workspace, is the most relevant to the current paper. Our work differs from [1] in that it does not require alignment of models with the real word prior to estimating pose, it employs more accurate pose initialization, is more resilient to mislocalized feature points, it can tolerate local minima in pose estimation and can be readily extended to multiple cameras.

3 Models and Features

Object models are a key ingredient of the proposed method. To obtain a complete, view independent model of an object, the latter has to be modelled using images from multiple viewpoints. Hence, each object is photographed individually in several images as a hand-held camera circumnavigates around it and then the acquired images are used to estimate the inter-frame camera motion and recover a corresponding 3D point cloud via SfM techniques [17]. An object model comprises of a set of 3D points from this point cloud, each accompanied by a SIFT image descriptor 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 neighbors and robustly fitting to them a plane. This procedure also identifies isolated 3D points that are filtered out from the final model.

During pose estimation, SIFT keypoints are detected in an image and then matched against those contained in an object model. The robustness of SIFT permits the reliable identification of features that have undergone large affine distortions between the image and the model. The established correspondences are used to associate the 2D image locations of feature locations with the 3D coordinates of their corresponding points on the objects surface. The procedure we initially evaluated for point matching used the standard ratio test for SIFT descriptor distances, as follows. Matches were identified by finding the two nearest neighbors to each descriptor from the image among those in the model, and only accepting a match if the distance to the closest neighbor was less than a fixed threshold of that to the second closest neighbor. This threshold can be adjusted to leniently establish more matches, or conservatively select the most reliable ones. It was observed experimentally that the ratio test yielded substantial proportions of erroneous matches. The F2P strategy from [10], also based on a ratio test, was also tested and found to be a viable choice in terms of the quality of produced matches, hence it was adopted as our matching technique.

Distances among SIFT descriptors are traditionally quantified with the Euclidean (L_2) norm. The SIFT descriptor is a weighted histogram of gradient orientations. Thus, irrespectively of the matching criterion, improvements in

matching are attained by substituting L_2 with histogram norms such as the Chi-squared (χ^2) distance [16]. Despite that other, more computationally demanding, distances such as the quadratic-Chi family or the circular Earth Movers were found to yield even better matching results, the χ^2 distance was eventually adopted as it offers the best performance / computational cost trade-off.

4 Pose Estimation

Pose estimation concerns determining the position and orientation of a camera given its intrinsics and a set of n correspondences between known 3D points and their image projections. This problem, also known as the Perspective-n-Point (PnP) problem, has received much attention due to is wide applicability in various domains. PnP is typically solved using non-iterative approaches that involve small, fixed-size sets of 3D-2D correspondences. For example, the basic case for triplets (P3P), was first studied in [4] whereas other solutions were later proposed in [2,8]. P3P is known to admit up to four different solutions, whereas in practice it usually has just two. As a result, a fourth point is used in practice for disambiguation. Minimal solutions to PnP are particularly important for estimating pose in a robust estimation framework, as the cardinality of each random sample is directly related to the total number of samples that need to be drawn in order to find a solution with acceptable confidence. On the other hand, being unable to combine more than the minimal number of correspondences, minimal solutions ignore much of the redundancy present in the data and hence suffer from inaccuracies. This is remedied by non-linear refinement, as follows.

4.1 Monocular Robust Pose Estimation with Non-linear Refinement

This section describes in more detail our approach for pose estimation in a single image. Starting with a set of 2D-3D point correspondences, a preliminary pose estimate is computed first and then refined iteratively. This is achieved by embedding a P3P solver into a RANSAC [2] framework that uses the MSAC redescending cost function for hypothesis scoring. Applied to the problem of pose estimation, RANSAC repetitively draws random quadruples of points and uses one triple with the P3P solver of [4] and the fourth point for verification to obtain a pose estimate. The best scoring pose hypothesis is retained as RANSAC's outcome and used to classify correspondences into inliers and outliers. By minimizing the reprojection error pertaining to all inliers, the pose computed by RANSAC is next refined to take into account more than three correspondences. Since it involves a non-linear objective function, this minimization is carried out iteratively with the Levenberg-Marquardt (L-M) algorithm, as explained next.

Denoting by $\mathbf{K}$ the 3×3 intrinsic calibration matrix and n corresponding 3D-2D points by $\mathbf{M}_i$ and $\mathbf{m}_i$, the pose computed with RANSAC is refined by using it as a starting point to minimize the cumulative image reprojection error

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

where $\mathbf{t}$ and $\mathbf{R}(\mathbf{r})$ are respectively the sought translation and rotation matrix parameterized using the Rodrigues rotation vector $\mathbf{r}$, $\mathbf{K} \cdot [\mathbf{R}(\mathbf{r}) \mid \mathbf{t}] \cdot \mathbf{M}_i$ is the predicted projection on the image of the homogeneous point $\mathbf{M}_i$ 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 Jacobians required by L-M were provided analytically using symbolic differentiation.

The minimization in (1) can be made more immune to noise caused by mislocalized image points by employing M-estimators [6]. The former substitute the squared-error of the residuals with a symmetric robust cost function $\rho()$ which increases less steeply than quadratically and/or down-weights points whose residual errors are too large. To ensure that $\rho()$ has a unique minimum at zero, it is common to choose it to be convex. However, non-convex cost functions are more effective in suppressing the influence of large errors at the cost of not guaranteeing uniqueness of minimum. An effective strategy is to start the process with a convex cost function, iterate until convergence, and then apply a few iterations with a non-convex one to eliminate the effect of large errors. Regardless of the exact form of the chosen cost function, it is stressed that M-estimation is robust to outliers due to mislocalization errors but not to false matches (i.e., gross errors which should be filtered out by other techniques like RANSAC prior to M-estimation). In our work, the application of M-estimators to pose refinement proceeds in two stages. The first stage employs the Fair convex cost function and the second Tukey's bi-weight for suppressing outliers.

4.2 Global Optimization for Pose Refinement

An issue with the objective function of (1) is that it is multimodal. Thus, nonlinear refinement with L-M initiated relatively far from the true minimum, runs the risk of getting trapped to a local minimum rather than converging to the true pose. To counter multiple minima, we have investigated the application of global optimization methods to pose estimation. A popular strategy for dealing with global optimization problems is to resort to multi-start procedures, which explore the feasible region by employing multiple runs of a local optimization algorithm started at several different points. More specifically, a multi-start algorithm selects a finite set of sample starting points from the feasible region. Local searches initiated at each of these points produce a set of local optima, the best of which is declared as the global optimum over the feasible region. Multi-start algorithms can make various choices of local search algorithms. Local search being a relatively expensive operation, multi-start methods seek to minimize the number of local searches performed. This is achieved by clustering together sampled points that lie in the region of attraction of the same local optimum. As a result, a single local search suffices for all points within the same cluster and yields considerable computational savings for the multi-start scheme. Among the various clustering methods available, the Multi Level Single Linkage (MLSL) algorithm [7] is one of the best, incorporating effective mechanisms for determining when to link sample points and when to terminate. SobolOpt [9] is an efficient MLSL variant that selects starting points with the aid of Sobol

sequences, which are pseudo-random low-discrepancy sequences that guarantee a good spatial distribution of samples. In this work, we have applied our implementation of SobolOpt to pose refinement, employing the L-M algorithm as its local search method. As a result, the pose estimation is rendered capable of escaping local minima and converging to the correct pose which would otherwise remain unreachable by plain L-M.

4.3 Binocular Pose Refinement

Estimating the pose as described in Sect. 4.1 employs a single camera. To improve accuracy with little additional overhead, a second viewpoint can be employed and the estimation can be extended to the binocular case by combining the reprojection error in two images. More specifically, assuming that two calibrated cameras are available, monocular pose estimation for each image is carried out as in Sect. 4.1. The pose of an object in one of the cameras (e.g. right) can be related to that in the other (i.e., left) with the aid of the extrinsic stereo calibration parameters. 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. Assuming a rigid stereo rig, $\mathbf{R_s}$ and $\mathbf{t_s}$ remain constant and can be estimated offline via extrinsic calibration. The binocular reprojection error consists of two additive terms, one for each image. Denoting the intrinsics for the left and right images by $\mathbf{K}^L$ and $\mathbf{K}^R$, the binocular reprojection error for n corresponding 2D-3D points in the left image and m in the right is defined as

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

where $\mathbf{t}$ and $\mathbf{R}(\mathbf{r})$ are the sought translation and rotation, $\mathbf{K}^L \cdot [\mathbf{R}(\mathbf{r}) \,|\, \mathbf{t}] \cdot \mathbf{M}_i$ is the projection of homogeneous point $\mathbf{M}_i$ in the left image, $\mathbf{K}^R \cdot [\mathbf{R}_s\mathbf{R}(\mathbf{r}) \,|\, \mathbf{R}_s\mathbf{t} + \mathbf{t}_s] \cdot \mathbf{M}_j$ is the projection of homogeneous point $\mathbf{M}_j$ in the right image and $\mathbf{m}_i^L$, $\mathbf{m}_j^R$ are the 2D points corresponding to $\mathbf{M}_i$ and $\mathbf{M}_j$ in the left and right images, respectively. It is noted that (2) circumvents the error-prone reconstruction of points via triangulation and does not limit the baseline of the two views nor calls for sparse feature or 3D point matching. It can also be extended to an arbitrary number of cameras. Similarly to the monocular case, a M-estimate of the reprojection error is minimized rather than its squared Euclidean norm. The minimization of (2) employs only the inliers of the two monocular estimations and can be started from the monocular pose computed for the left camera. Since this is assumed to be already close to the true minimum, application of the global optimization scheme of Sect. 4.2 is not essential. Nevertheless, 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 is erroneous, 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 as the most accurate one.

5 Object Detection and Recognition

Assume that a set of models corresponding to known objects and a strongly calibrated binocular camera system are available. SIFT features and their corresponding descriptors are extracted independently from each image. Optionally, and in order to increase efficiency, feature detection can be restricted only within certain regions of interest (ROIs) in images. Such regions are intended as a means for directing the system's attention to where objects might approximately be and need not be very accurate or in 1-1 correspondence with the objects actually contained in an image. In this work, ROIs are determined by a color-based foreground extraction process [20]. SIFT descriptors are matched against those of the models to establish putative correspondences. For each model giving rise to sufficiently many such correspondences, a hypothesis concerning the presence of the corresponding object in the image is formed. Such hypotheses are tested by using the 3D coordinates of matched model points to estimate object pose as detailed in Sections 4.1 & 4.2 and verify that the model accounts for the observed arrangement of features with adequate confidence. Confidence is quantified by considering the proportion c_p and number c_n of pose estimation inliers supporting a hypothesis and considering an object to be present if $c_p > 90\%$ and $c_n > 40$. A by-product of this process is an estimate of the pose of each detected object. Evaluation of all hypotheses for an image identifies all objects in it. Being independent of each other, the evaluation of each hypothesis in our implementation runs in parallel with the rest, using different CPU cores. Applied to each image, the aforementioned process yields two sets S_L and S_R comprised of the objects found in each. To increase accuracy, the system strives to perform pose estimation using both cameras for objects in the intersection of S_L and S_R, using the approach of Sect. 4.3. For the remaining objects found only in one image, their monocular poses are employed.

6 Experiments

Two groups of experiments were performed. The first employed synthetic images generated with poses known beforehand. The second group utilized real images for which the true poses were inferred via careful object placement on a calibration grid. To generate synthetic images, a textured mesh model of an object was obtained in addition to its sparse keypoint model recovered as in Sect. 3. Utilization of a mesh enables dense rendering of the object while taking into account its self-occlusions. The mesh was obtained using the ARC3D service [19] and was aligned with the sparse keypoint model. Finally, a conventional OpenGL renderer was employed to render the mesh at selected poses against a black background. The object rendered for the experiment was the green parallelepiped of Fig. 1 with size $45 \times 45 \times 90\, mm^3$. The virtual camera orbited it in 60 frames with its optical axis oriented towards the object's centroid. The experiment was conducted in four conditions, in which the radius of the circular trajectory was modulated at 500, 750, 1000 and $1500\, mm$ from the object's centroid. The camera moved

Table 1. Mean and standard deviation for the translational ($\mathbf{t}$, mm) and rotational ($\mathbf{R}$, $^\circ$) pose errors for methods (a), (b), and (c)

Radius	$\mathbf{t}_a$	$\mathbf{R}_a$	$\mathbf{t}_b$	$\mathbf{R}_b$	$\mathbf{t}_c$	$\mathbf{R}_c$
$500\,mm$	9.40 (7.33)	0.38 (0.25)	9.01 (5.85)	0.34 (0.20)	6.56 (4.21)	0.26 (0.17)
$750\,mm$	15.02 (12.82)	0.58 (0.41)	10.42 (7.31)	0.43 (0.27)	8.20 (5.23)	0.34 (0.18)
$1000\,mm$	17.07 (13.71)	0.70 (0.45)	11.63 (9.68)	0.48 (0.43)	9.05 (6.36)	0.40 (0.23)
$1500\,mm$	16.96 (12.24)	0.73 (0.49)	13.09 (11.57)	0.50 (0.39)	10.65 (7.22)	0.48 (0.26)

on a plane that was $185\,mm$ above that upon which the object was placed. The ground truth pose was compared against the estimates obtained from three methods: (a) monocular pose estimation (Sect. 4.1 and 4.2), (b) binocular pose estimation using 3D points reconstructed with stereo triangulation followed by absolute orientation [5], and (c) binocular pose estimation using the joint reprojection error of points in two views (Sect. 4.3). Only one image was employed for each application of method (a). Methods (b) and (c) employed a binocular pair comprised of the same image as in (a) and a second image which was four frames ahead of the first. Thereby, the two cameras are verging at the object at a relative angle of 24°. Table 1 summarizes the mean and standard deviation of the error pertaining to the translational and rotational components of the pose estimates for the four conditions of the experiment. The translational error between an estimate $\widehat{\mathbf{t}}$ and the true translation $\mathbf{t}$ is computed as $||\mathbf{t} - \widehat{\mathbf{t}}||$, whereas the rotational error between $\widehat{\mathbf{R}}$ and $\mathbf{R}$ is $\arccos((\mathrm{trace}(\mathbf{R}^{-1}\widehat{\mathbf{R}}) - 1)/2)$ and corresponds to the amount of rotation about a unit vector that transfers $\mathbf{R}$ to $\widehat{\mathbf{R}}$. Clearly, method (c) outperforms the other two in all conditions and, expectedly, error grows with the camera distance from the target. Both binocular methods (b and c) outperform the monocular one (a), as expected.

The binocular method of Sect. 4.3 is evaluated next with the aid of real images. To obtain ground truth for object poses, a checkerboard was used to guide the placement of a target 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. Note that these presumed locations include minute calibration inaccuracies as well as human errors in object placement. The green object of Fig. 1 was placed and aligned upon every checker of the 8×12 checkerboard visible 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 960×1280 pixels, and its FOV was $16^\circ \times 21^\circ$. The mean translational error in these 96 trials was $2.4\,mm$ with a deviation of $1.5\,mm$. Total running time per frame was around $600\,ms$ on an Intel Core i7 CPU 950 @ 3.07GHz.

The accuracy of pose estimation was also evaluated in the presence of modulated occlusions. In this experiment, objects have been occluded in both images of a binocular pair by applying an increasingly larger mask to each bounding box of an object. This mask was expanded from the bottom of the bounding box

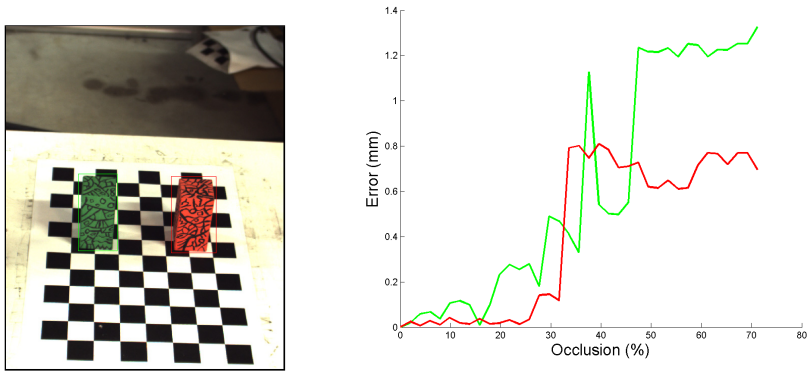

Fig. 1. Estimation error in the presence of increasing occlusions. Left: an image from a binocular pair (courtesy of Profactor GmbH). Right: translational components of pose estimation errors, plotted as functions of the occlusion percentage for the two objects.

Fig. 2. Pose estimation under occlusions. Left: an image from a binocular pair. Right: point clouds transformed with their estimated poses to the checkerboard coord. frame.

towards its top, masking out object pixels. The experiment was performed for both objects of Fig. 1. We observe that the method is quite robust to occlusions, as it provides less than 5 mm of translational error wrt the unoccluded pose for occlusions up to 70% in the images. In another qualitative experiment, we employed six objects severely occluding each other. In addition, a cardboard box was also placed in front of the objects (see Fig. 2). Despite the severity of occlusions, top faces of objects provided sufficiently many features to the binocular method which, as shown in the right figure, yields fairly accurate pose estimates.

7 Conclusion

The paper has presented an approach for rigid object detection and pose estimation that was shown experimentally to be very accurate and robust to occlusions. Future work will address the incorporation of geometric constraints in the feature matching process and the extension of the method to track moving objects.

References

1. Collet Romea, A., Berenson, D., Srinivasa, S., Ferguson, D.: Object Recognition and Full Pose Registration from a Single Image for Robotic Manipulation. In: Proc. of ICRA 2009 (May 2009)
2. Fischler, M., Bolles, R.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. CACM 24, 381–395 (1981)
3. Gordon, I., Lowe, D.G.: What and Where: 3D Object Recognition with Accurate Pose. In: Ponce, J., Hebert, M., Schmid, C., Zisserman, A. (eds.) Toward Category-Level Object Recognition. LNCS, vol. 4170, pp. 67–82. Springer, Heidelberg (2006)
4. Grunert, J.: Das pothenotische Problem in erweiterter Gestalt nebst über seine Anwendungen in Geodäsie. Grunerts Archiv für Mathematik und Physik (1841)
5. Horn, B.: Closed-Form Solution of Absolute Orientation Using Unit Quaternions. J. Optical Soc. Am. A 4(4), 629–642 (1987)
6. Huber, P.: Robust Statistics. Wiley (1981)
7. Kan, A.R., Timmer, G.: Stochastic Global Optimization Methods, Part I & II (Clustering Methods & Multi-Level Methods). Math. Program. 39(1), 27–78 (1987)
8. 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)
9. Kucherenko, S., Sytsko, Y.: Application of Deterministic Low-Discrepancy Sequences to Nonlinear Global Optimization Problems. Comput. Optim. Appl. 30(3), 297–318 (2005)
10. 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)
11. Lowe, D.: Three-Dimensional Object Recognition from Single Two-Dimensional Images. Artificial Intelligence 31(3), 355–395 (1987)
12. Lowe, D.: Fitting Parameterized Three-Dimensional Models to Images. IEEE Trans. Pattern Anal. Mach. Intell. 13(5), 441–450 (1991)
13. Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vis. 60(2), 91–110 (2004)
14. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A Comparison of Affine Region Detectors. Int. J. Comput. Vis. 65(1-2), 43–72 (2005)
15. Rothganger, F., Lazebnik, S., Schmid, C., Ponce, J.: 3D Object Modeling and Recognition Using Local Affine-Invariant Image Descriptors and Multi-View Spatial Constraints. Int. J. Comput. Vis. 66(3), 231–259 (2006)
16. 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)
17. Snavely, N., Seitz, S., Szeliski, R.: Photo Tourism: Exploring Photo Collections in 3D. ACM Trans. Graph. 25(3), 835–846 (2006)
18. Vacchetti, L., Lepetit, V., Fua, P.: Stable Real-Time 3D Tracking Using Online and Offline Information. IEEE Trans. Pattern Anal. Mach. Intell. 26(10), 1385–1391 (2004)
19. Vergauwen, M., Gool, L.V.: Web-based 3D Reconstruction Service. Mach. Vision Appl. 17(6), 411–426 (2006)
20. Zivkovic, Z.: Improved Adaptive Gaussian Mixture Model for Background Subtraction. In: Proc. of ICPR, vol. (2), pp. 28–31 (2004)