Simultaneous Pose, Focal Length and 2D-to-3D Correspondences from Noisy Observations

Adrian Penate-Sanchez
apenate@iri.upc.edu
Eduard Serradell
eserradell@iri.upc.edu
Juan Andrade-Cetto
cetto@iri.upc.edu
Francesc Moreno-Noguer
fmoreno@iri.upc.edu

Institut de Robòtica i Informàtica
Industrial
CSIC-UPC
08028, Barcelona, Spain

Abstract

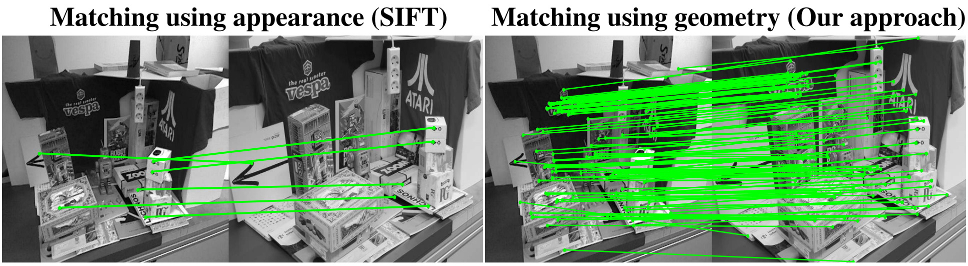
Simultaneously recovering the camera pose and correspondences between a set of 2D-image and 3D-model points is a difficult problem, especially when the 2D-3D matches cannot be established based on appearance only. The problem becomes even more challenging when input images are acquired with an uncalibrated camera with varying zoom, which yields strong ambiguities between translation and focal length. We present a solution to this problem using only geometrical information. Our approach owes its robustness to an initial stage in which the joint pose and focal length solution space is split into several Gaussian regions. At runtime, each of these regions is explored using an hypothesize-and-test approach, in which the potential number of 2D-3D matches is progressively reduced using informed search through Kalman updates, iteratively refining the pose and focal length parameters. The technique is exhaustive but efficient, significantly improving previous methods in terms of robustness to outliers and noise.

1 Introduction

Estimating the 3D pose of a camera with respect to a 3D object typically requires to know in advance the calibration parameters of the camera and establishing a set of 3D-to-2D point correspondences between a 3D model of that object and an input image. If a reference image is registered to the 3D model, usually, the correspondence problem becomes a 2D-to-2D one. In these cases by using point descriptors, like SIFT [10], and by performing robust matching algorithms, such as [3, 6], we eliminate outliers and are able to calculate the camera pose.

Matching methods based on RANSAC rely on having a small outlier rate. This way, the probability of obtaining a correct minimal set of points is high enough to be achieved in

This work has been partially funded by Spanish Ministry of Economy and Competitiveness under project PAU+ DPI2011-27510; by the EU project ARCAS FP7-ICT-2011-28761 and by the ERA-Net CHISTERA project VISEN PCIN-2013-047.


Matching using appearance (SIFT)

Matching using geometry (Our approach)

Figure 1: Inlier correspondences of a matching algorithm that uses appearance information (left) and our matching algorithm (right) that only uses the 3D and 2D location of the points to search for the correspondences. Matching between the intensity component of a Kinect camera and a Canon EOS 60D camera, together with the viewpoint changes and the self-occlusions, differences in terms of image noise and resolution jeopardize appearance matching producing a very small set of correct matches. In contrast, our method based purely on geometric information is able to retrieve a much larger number of correspondences, accurately computing the relative pose and focal length under such conditions.

fixed time. If the percentage of outliers increases, the number of RANSAC loops needed to obtain a correct set of minimal points grows exponentially. Also, as shown in [15], using the minimal set of points might not yield the best pose if there is noisy data in the system. The percentage of outliers depends on several factors: the precision recall of the point descriptor, changes in appearance, changes in points of view, repeated patterns, self occlusions, etc. We propose a matching algorithm that performs robust matching under the presence of a large percentage of outliers. Our approach is a generalization of the work in [12] to deal with uncalibrated cameras and noisy 3D information. We show experiments using 3D data obtained with a Kinect camera [21] as an example of the kind of contexts we solve.

To tackle all these issues we propose to split the initial prior distribution of the combined pose and focal length estimate into an arbitrary large number of Gaussian priors. These priors are spread within very rough bounds of where the pose and focal length are expected to be. At runtime, each of these priors is used to guide the search for the 3D-to-2D correspondences, while progressively pruning the number of potential candidate matches and refining the pose and focal length values. Repeating this process for each of the priors guarantees an exhaustive exploration of the solution space at a limited computational cost.

Experiments in both synthetic and real data will show that the proposed approach is robust to large levels of 2D and 3D noise and clutter, yielding reasonable results for outlier rates of up to 80%. This significantly outperforms competing approaches [2, 3, 9], and is comparable to methods that assume a calibrated camera [12].

2 Related Work

Pose estimation techniques that maximize image similarity, such as [1], are not applicable in our context due to their limited ability to deal with significant differences in appearance and reduced capture range, as that of Fig. 1. We shall therefore consider only techniques that explicitly perform matching using the geometric structure of the 2D and 3D point sets.

The robust estimation of correspondences between two sets of points has been historically solved by hypothesize and test algorithms such as RANSAC [6] and Least Median

Squares [17]. They rely on a random sampling of minimal subsets to generate model hypotheses, and favor the one that best explains most of the data points. Unfortunately, in these methods, computational complexity scales exponentially with the number of model parameters and the size of the point set. Among the several variations of the original RANSAC algorithm, Guided-MLESAC [23] and PROSAC [3] avoid sampling unlikely correspondences by using appearance based scores and thus are not applicable to our problem. Similarly, GroupSAC [13] uses image segmentation to sample more efficiently the data. Other techniques of the same family such as Preemptive RANSAC [14] or ARRSAC [16] work within a limited time scenario thus increasing the probability of not reaching the best estimate. Finally, MultiGS [2] accelerates the search strategy by guiding the sampling with information from residual sorting and is able to account for multiple structures appearing in the scene. In the experimental section, we have chosen to compare our approach against this method which we consider representative of the whole family.

In the absence of robust appearance information graph matching can be used, as proposed by [5]. Yet, due to its high computational cost, these methods are only applicable to small graphs. While such approaches allow global optimization, they cannot be used with large intra-image distances due to very different points of view of the scene, or when the number of outliers is excessive, which are the cases we consider in this paper.

The L_∞ technique proposed in [9], uses second-order cone programming, and guarantees optimality under the L_∞ norm, for different geometric structure and motion problems, including the camera pose estimation considered in this work. However, this particular metric is highly sensitive to outliers, as pointed out in [8]. Even when it is possible to address in part the outlier removal problem as proposed in [22], the L_∞ solution for the camera pose estimation only performs as well as the standard L_2 norm.

Other approaches simultaneously solve for pose and correspondences purely from geometric point matching. Of these, SoftPOSIT [4] uses an iterative technique to generate correspondence candidates, but the global minimum can not be guaranteed. Our approach is inspired in the Blind PnP algorithm [12], where local optimality is alleviated introducing the scene geometry as pose priors, modeled as a Gaussian mixture model, and progressively refined by hypothesizing correspondences. Incorporating each new candidate in a Kalman filter rapidly reduces the number of potential 2D matches for each 3D point and makes it possible to search the pose space sufficiently fast for the method to be practical. More recent techniques [19] use robust estimation in a final stage to refine the pose. Unfortunately, the approach cannot be applied straightforward to the uncalibrated case due to the ambiguities between focal length and the pose translation vector.

3 Algorithm

The structure of the algorithm we propose is similar in spirit to the Blind PnP algorithm [12]. In an initial stage we split the solution space into a set of Gaussian clusters. Then, we progressively explore each of these clusters to simultaneously establish the 2D-to-3D correspondences and refine the camera pose and focal length. In contrast to the original Blind PnP formulation, we are considering an uncalibrated camera, with the focal length as an additional parameter to estimate. This yields and extra degree of complexity to the problem, as there are large ambiguities between changes of the camera focal length and translations along the optical axis. In addition, the proposed approach takes into account the uncertainties in the 3D model, characteristic of range sensors such as ToF or Kinect cameras [7].

3.1 Problem Formulation

Let us assume we are given a reference model made up of M 3D points $\mathcal{X} = \{\mathbf{x}_i\}$, with their 2D correspondences on a reference image, and a set of N 2D points $\mathcal{U} = \{\mathbf{u}_j\}$ on an input image, acquired with an uncalibrated camera. We consider that correspondences between the 2D and 3D sets is possible since both point sets were extracted using the same interest point detector over the reference and input images. Let us denote by $\mathbf{p}$ a 6-dimensional vector, parameterizing the rotation and translation that aligns the camera with respect to the 3D point set coordinate system, and let f be the unknown camera focal length. We additionally assume a camera with square pixels, and with the principal point located at the center of the image, and hence, being the focal length the only unknown intrinsic parameter. Our goal is to retrieve the pose of the camera and its focal length. As the 3D-to-2D correspondences are unknown, they need to be retrieved together with the pose and focal length parameters. This can be formulated as an optimization problem, retrieving $\mathbf{p}$ and f such that the reprojection error between the projected 3D points $\mathbf{x}_i$ and their corresponding matches $\mathbf{u}_j$ is minimized,

$$\underset{\mathbf{p},f}{\text{minimize}} \sum_{i=1}^{M} \mathsf{Inlier}(\|\mathsf{Proj}(\mathbf{x}_i;\mathbf{p},f) - \mathsf{Match}(\mathbf{x}_i;\mathcal{U})\|) \tag{1}$$

where $\mathsf{Proj}(\mathbf{x}_i;\mathbf{p},f)$ returns the 2D perspective projection $\tilde{\mathbf{u}}_i$ of a 3D point $\mathbf{x}_i$ given the pose and focal length parameters; $\mathsf{Match}(\mathbf{x}_i;\mathcal{U})$ returns the $\mathbf{u}_j \in \mathcal{U}$ that is closest to $\tilde{\mathbf{u}}_i$; and

$$\mathsf{Inlier}(d) \begin{cases} d & \text{if } d < Max\_distance\_inlier \\ \mathsf{Penalty\_outlier} & \text{otherwise} \end{cases}$$

is a function that penalizes points whose reprojection error is above a $Max\_distance\_inlier$ threshold . This is to avoid local minima, otherwise, small sets of candidate matches can get a lower error than the actual solution.

3.2 Modeling Uncertainty

A straight-forward way to minimize Eq. 1 would be to use a RANSAC-like approach [6], and repetitively hypothesize sets of four 3D-to-2D correspondences until one of them yields an estimate of $\mathbf{p}$ and f that brings the reprojection error below a certain threshold. Unfortunately, since these methods do not introduce constraints on the potential set of 2D candidates that can match each 3D point, they are only computationally tractable for a relatively small number of features. In order to make the minimization of Eq. 1 tractable, we follow a similar strategy as in [12], and split the solution space into an arbitrary large number of Gaussian regions. At runtime, each of these regions is explored in turn, guiding a matching process for each 3D point. By splitting the search space in these small regions, the total number of potential 2D candidates for each 3D point is significantly reduced.

Estimation of the pose and focal length priors is done in a pre-processing stage. We first define the bounds of these parameters. For the pose, we acquire several images of the 3D object at the extremal positions and orientations of the working space. This is used to build an hyper-box in the 6-dimensional pose space, which is then subsampled using Montecarlo. Expectation Maximization is run over these samples to compute the N_p Gaussian priors on the pose, defined by a set of mean poses $\mathbf{p}_k$, $k = 1,\ldots,N_p$, and a set of 6×6 covariance matrices $\Sigma_k^{\mathbf{p}}$. Similarly, the range of feasible focal lengths, is split into N_f Gaussian priors, defined by mean values f_l and the corresponding one-dimensional variances σ_l^f, $l = 1,\ldots,N_f$.

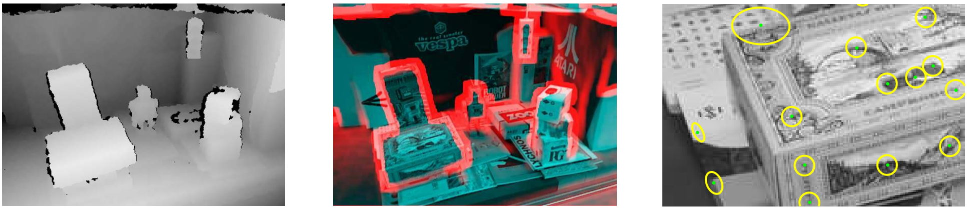


Figure 2: **Uncertain 3D model**. **Left:** 3D model acquired with a Kinect camera. Regions in which the 3D data is most uncertain are depth discontinuities. **Center:** We detect the uncertain regions –shown in red– computing depth covariances within local neighborhoods. **Right:** A 3D covariance is assigned to each 3D model point and propagated to the image plane. This is used to limit the area where to search for potential match candidates.

One of the key ingredients of our approach is that it lets us handle uncertain 3D models, such as those obtained from a Kinect camera. It is well known that these sensors suffer from inaccuracies, especially at depth discontinuities (see Fig. 2). In order to inject this uncertainty into the optimization process, each 3D model point $\mathbf{x}_i$ is assigned a covariance $\Sigma_i^{\mathbf{x}}$, computed considering the depth variations of their neighboring points. During the optimization process, those points with larger uncertainties will have smaller impact in the computation of the solution. We also assign an uncertainty $\Sigma_j^{\mathbf{u}}$ to each 2D measurement.

3.3 Optimization

Given the sets $\mathcal{X}$ and $\mathcal{U}$, the pose and focal length priors, and the 3D and 2D uncertainties, we proceed to the optimization of Eq. 1 by progressively exploring each pair of priors $\{\mathbf{p}_k, \Sigma_k^{\mathbf{p}}\}; \{f_l, \Sigma_l^f\}$, using the following steps:

3.3.1 Projecting uncertainties onto the image plane

To limit the number of potential 2D match candidates for each 3D point $\mathbf{x}_i$, we project them onto the image plane and compute the uncertainty in the projection assuming independent contribution from all three sources: 3D point uncertainty, pose uncertainty, and focal length uncertainty. The result is a Gaussian distribution with mean $\tilde{\mathbf{u}}_i$ and covariance $\Sigma_i^{\tilde{\mathbf{u}}}$:

$$\begin{aligned} \tilde{\mathbf{u}}_i &= \mathsf{Proj}(\mathbf{x}_i; \mathbf{p}_k, f_l) \\ \Sigma_i^{\tilde{\mathbf{u}}} &= \mathbf{J}_{\mathbf{x}} \Sigma_i^{\mathbf{x}} \mathbf{J}_{\mathbf{x}}^\top + \mathbf{J}_{\mathbf{p}} \Sigma_k^{\mathbf{p}} \mathbf{J}_{\mathbf{p}}^\top + \mathbf{J}_f \sigma_l^f \mathbf{J}_f^\top \ , \end{aligned} \tag{2}$$

where $\mathbf{J}_g = \frac{\partial \mathsf{Proj}(\mathbf{x}_i; \mathbf{p}_k, f_l)}{\partial g}$ is the Jacobian of the projection function with respect to each of the uncertain parameters $g = \{\mathbf{x}, \mathbf{p}, f\}$. Using the Gaussian distribution $\{\tilde{\mathbf{u}}_i, \Sigma_i^{\tilde{\mathbf{u}}}\}$, we can define a search region for the point $\mathbf{x}_i$, and consider as potential candidates $\mathcal{PC}(\mathbf{x}_i)$ all points $\mathbf{u}_j \in \mathcal{U}$ whose Mahalanobis distance is below a threshold $\mathsf{Max\_Mah}$, i.e:

$$\mathcal{PC}(\mathbf{x}_i) = \left\{ \mathbf{u}_j \in \mathcal{U} \ \text{ s.t. } \ (\mathbf{u}_j - \tilde{\mathbf{u}}_i)^\top (\Sigma_i^{\tilde{\mathbf{u}}})^{-1} (\mathbf{u}_j - \tilde{\mathbf{u}}_i) < \mathsf{Max\_Mah}^2 \right\} \cup \{\emptyset\} \tag{3}$$

where $\emptyset$ denotes the possibility that $\mathbf{x}_i$ is in fact an outlier and does not have a 2D image correspondence.

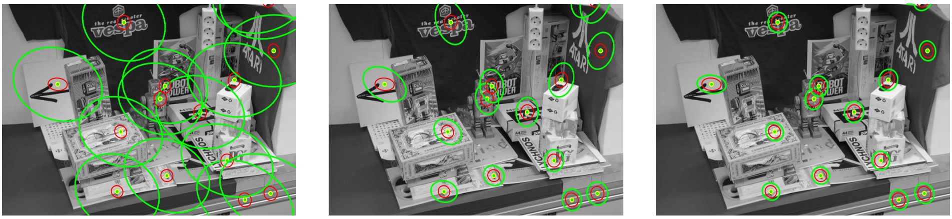


Figure 3: **Limiting the number of potential candidates. Left:** Search region obtained after projecting the three terms of Eq. 2 independently. **Center and Right:** Refinement of the search space, after establishing correspondences.

3.3.2 Local Guided Hypothesize-and-Test

Once we have defined the set of potential 2D candidates for all the 3D points, we start a hypothesize and test strategy, similar to what is done in a standard RANSAC algorithm. Yet, in contrast to RANSAC we only need to establish potential matches within local neighborhoods. In addition, after a hypothesis has been made, we use a Kalman filter formulation to shrink the size of the Gaussian regions associated to the pose and focal length, to further reduce and guide the set of potential candidates in each iteration Figure 3. We initialize this step choosing the least ambiguous point

$$\mathbf{x}_i^* = \underset{\mathbf{x}_i \in \mathcal{X}}{\arg\min} \left|\mathcal{PC}(\mathbf{x}_i)\right| , \tag{4}$$

i.e, the 3D point with the lowest number of potential candidates. In doing so we start with a 3D point with low uncertainty, since these are the ones with smaller search regions for potential matches, in the Mahalanobis sense. We then hypothesize the match $\{\mathbf{x}_i^*, \mathbf{u}_j^*\}$, where $\mathbf{u}_j^*$ is the 2D candidate within $\mathcal{PC}(\mathbf{x}_i^*)$ that is closest to $\tilde{\mathbf{u}}_i^*$ in terms of Mahalanobis distance. We then use standard Kalman filter equations to update the pose and focal length and reduce their associated covariances:

$$\begin{array}{ll} \mathbf{p}_k^+ = \mathbf{p}_k + \mathbf{K}_\mathbf{p}(\mathbf{u}_j^* - \tilde{\mathbf{u}}_i^*) & \quad f_l^+ = f_l + \mathbf{K}_f(\mathbf{u}_j^* - \tilde{\mathbf{u}}_i^*) \\ \Sigma_k^{\mathbf{p},+} = (\mathbf{I} - \mathbf{K}_\mathbf{p}\mathbf{J}_\mathbf{p})\Sigma_k^\mathbf{p} & \quad \sigma_l^{f,+} = (1 - \mathbf{K}_f\mathbf{J}_f)\sigma_l^f \end{array} .$$

The new pose, focal length and covariance matrices are used to project again the 3D points onto the image, and define new and smaller search regions.

3.3.3 Backtracking and Iterating over all Priors

Maintaining the strategy of choosing the 3D point with less potential matches, and the 2D point closest to its projection, we hypothesize new 3D-to-2D matches and refine the pose and focal length posteriors and their associated uncertainties. This process is repeated until the Kalman update terms become negligible, usually in less than five iterations. Upon convergence, we project the remaining 3D points onto the image and match them to the nearest 2D feature point. 3D points whose nearest neighbor distance is larger than *Max_distance_inlier* are classified as outliers. Using both the inlier and outliers points, we compute the error of Eq. 1 and stop the algorithm for the current prior set $\{\mathbf{p}_k, \Sigma_k^\mathbf{p}\}; \{f_l, \Sigma_l^f\}$ if the error falls below a given threshold. If not, we backtrack through the list of 3D-to-2D matches to change the assignments and repeat the guided search and refinement process. When no more assignments are available, we repeat the process with a different pose and focal length prior.

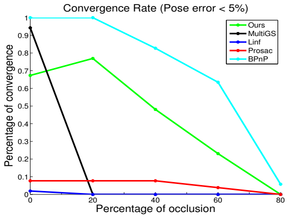


Figure 4: Convergence rates obtained for increasing levels of occlusion. The solution is considered correct if both the relative errors between the estimated pose and focal length, and their ground truth values do not exceed 5, 10 and 20% (from left to right).

3.3.4 Accelerating the Optimization and Final Refinement

We can further increase the efficiency of the algorithm if a wrong pose or focal length prior is detected before completing the exploration over all corresponding points. One criteria for doing this is setting a threshold to the maximum number of outliers allowed. Following a branch and bound strategy, we terminate the exploration for a specific pose and focal length prior when the minimum number of outliers reaches a specific threshold. Being conservative, we have set this threshold to $0.8M$, i.e, we accept a maximum of 80% of outliers.

Finally, at the termination of the search, our algorithm yields a set of correspondences and estimations of the focal length and pose parameters. We further refine these results by performing a final RANSAC-based step, in order to eliminate residual mismatches. Note however, that this approach does not need to establish matches, but just remove a very reduced number of incorrect ones, and thus, its computational cost is negligible.

4 Results

Evaluation has been done on synthetic and real data. Synthetic results accurately evaluate our approach in comparison to state-of-the-art. Real data results provide a qualitative assessment of the method in a challenging scenario with high levels of clutter and noise.

4.1 Synthetic Results

We now present results on synthetic data by evaluating our algorithm in controlled experiments with known ground truth. For these experiments we compare our approach, denoted Uncalibrated BlindPnP (UBPnP), with Multi Guided Sampling (MultiGS) [2], the L_∞ method (Linf) [9] and with PROSAC [3]. We also compare it against BlindPnP [12] for which we inject the true calibration parameters, and use the results as a baseline to give significance to the reported accuracy estimates. We synthetically generated the 3D model by randomly sampling $N = 50$ points from a cube of dimensions $x \in [-1,1]$, $y \in [-1,1]$, $z \in [-1,1]$, and selecting the ground truth camera pose from inside a torus surrounding the point set. The camera optical axes are chosen randomly to point anywhere on the 3D model. Then, the 2D points are produced by projecting the model onto a 640 × 480 image, using a calibration matrix where we allowed the focal length to vary within the interval $f = [600 - 1200]$. We performed 50 trials on each of the different setups with increasing percentages of occlusion $p_o \in \{0, 20, \ldots, 80\}$. Pose clusters are created with a 30-component Gaussian Mixture Model, with 10 additional components to cluster the focal length range.

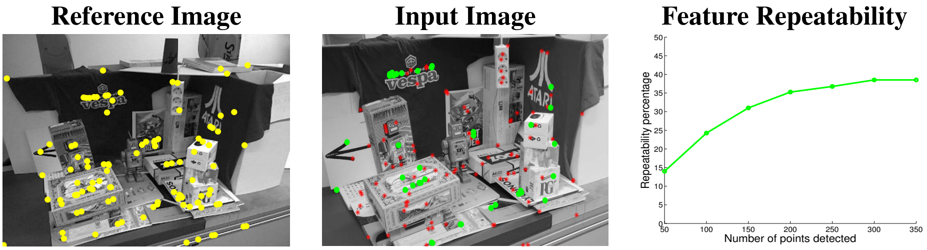


Figure 5: **Feature repeatability.** The image in the middle shows in green the feature points that have been consistently detected in both the reference and input images, and in red, the features that have only been detected in the input image. The graph on the right shows how the percentage of repeated features (potential matches) increases with the number of detected points. In our experiments we work with 50-100 model points, and hence, we need to be robust to percentages of up to $75-85\%$ of outliers.

The standard situation in which Linf, MultiGS and PROSAC are used, is one in which a set of potential matches, computed using texture information, are given in advance and their goal is to reject mismatches. In our case, since the texture is not available, we calculate the potential matches for these algorithms using only the pose and focal length clusters in which the correct solution lies. We have projected the covariances to get all the matching candidates for each 3D point and kept all the possible pairs as candidate matches. In addition, for PROSAC we replaced its score function based on appearance similarity, by a similarity function defined by the Euclidean distance between the projected point and each potential 2D candidate. Note that this criterion should be informative enough considering the points were projected from the correct pose prior.

Results comparing the convergence ratio in all the experiments are shown in Fig.4. This convergence ratio, represents the percentage of experiments for which the relative rotation, translation, and focal length errors are below a certain threshold (5%, 10% and 20%). These thresholds are chosen to reflect the fact that the reconstruction error is more sensitive to a correct estimate of the rotation than to a correct estimate of the translation and focal length terms. Note that the performance of our approach breaks only once occlusion levels larger than 50% are reached, being very similar to the calibrated BlindPnP. As seen in the charts, the methods built to rely on appearance, even with the given advantages, cannot perform as well as purely geometric methods when appearance cannot be used. Regarding computational time, UBPnP scales linearly the complexity of BPnP as a result of introducing an extra dimension, clustered using additional N_f Gaussian priors. This means a computation time between 10 and 1000 seconds for increasing sizes of the point set going from 20 to 80 points.

4.2 Results with Real Data

The technique has been tested for the registration of imagery acquired with a Canon EOS 60D camera varying the pose and focal length, to a 3D model acquired with a Kinect camera. The Kinect intensity image was used as a reference to extract the model points. Points of interest were extracted using *DoG*. $M = 100$ model points were computed on the reference image, and between $N = 100-150$ were computed on each of the 50 input images.

One important issue we had to handle is that of obtaining a minimum number of keypoints that consistently appear in the reference and input images. This problem is especially

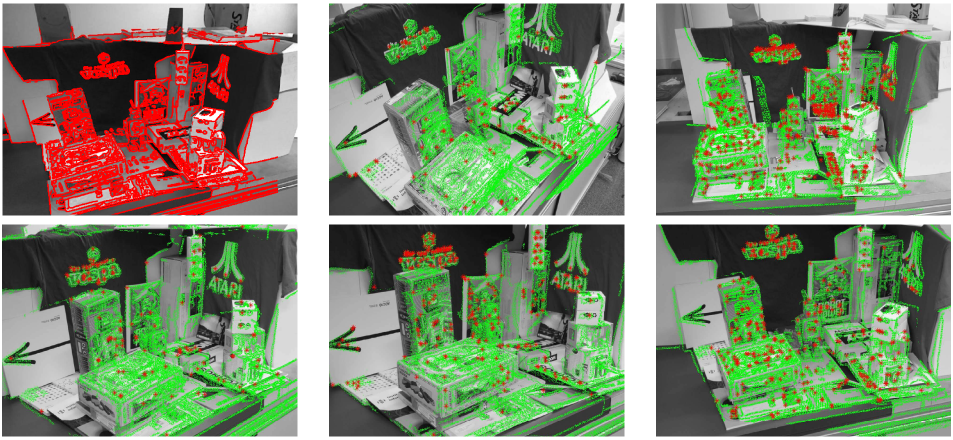


Figure 6: Real experiments. **Top Left:** Reference image registered to the 3D model. **Others:** Reprojection of the 3D model onto the input images after estimating pose, focal length and 3D-to-2D matches with our approach.

critical in our framework, as our algorithm (and also the competing approaches) can only handle, in a reasonable amount of time, sets of about 100 points. As shown in Fig. 5 the percentage of inliers decreases with the number of detected points, dropping from levels of about a 40% of inliers for 350 models points, to 25% for 100 points, which is the number of 3D model points we used in our experiments. This is therefore a challenging scenario to test the robustness of our approach in the presence of outliers.

To compute the pose priors we acquired several images at extremal positions and orientations of the working space and manually registered them with the 3D map. The poses of these images were used as bounds for fitting $N_p = 100$ pose priors. The $18-150$mm range of the lens was split into $N_f = 10$ Gaussian intervals. Given that the ground truth pose of the query images is not available, we evaluated the method according to the minimization of the reprojection error. Fig. 6 shows several images in which the boundaries of the 3D model are reprojected after computing the correspondences, pose and focal length. Even dealing with large differences in viewpoint, the reprojection results are very accurate.

5 Conclusion

Simultaneously estimating the camera position, orientation, focal length and establishing 3D-to-2D correspondences between model and image points, poses a challenging optimization problem which can hardly be solved without prior information. Most current approaches rely on appearance information to first solve the correspondences and then retrieve the pose and focal length while rejecting missmatches. Yet, there are many situations in which the appearance is either not available or not a reliable cue.

In the absence of appearance, we propose to use only geometric priors, which are just rough approximations of the pose and focal length solution space. By progressively exploring these priors we are able to efficiently prune the potential number of 3D-to-2D matches, while reducing the uncertainty of the pose and focal length estimates. The method is shown to be highly resilient to clutter and noise on the image features and in the 3D model. The latter is especially suited for dealing with 3D models obtained from noisy range sensors, such as the Kinect or Time of Flight cameras.

Regarding future work, we plan to integrate our framework in a setting in which generic appearance models can be aggregated for the same keypoint as observed from multiple vantage points. Along these lines we will take advantage of recent descriptors such as [24] or the DaLI [11]. The latter, is specially interesting because it will let us bring our approach from a rigid to a deformable domain, like in [18, 20], but without the strong assumption of having to know the camera calibration parameters in advance.

References

[1] M. Calonder, V. Lepetit, M. Özuysal, T. Trzcinski, C. Strecha, and P. Fua. Brief: Computing a local binary descriptor very fast. *IEEE Transactions on Pattern Analylis and Machine Intelligence*, 34(3):1281–1298, 2012.

[2] T.J. Chin, J. Yu, and D. Suter. Accelerated hypothesis generation for multi-structure robust fitting. In *European Conference on Computer Vision, ECCV*, volume 6315 of *Lecture Notes in Computer Science*, pages 533–546. Springer, 2010.

[3] O. Chum and J. Matas. Matching with PROSAC - progressive sample consensus. In *Conference on Computer Vision and Pattern Recognition, CVPR*, pages 220–226. IEEE, 2005.

[4] P. David, D. DeMenthon, R. Duraiswami, and H. Samet. SoftPOSIT: Simultaneous pose and correspondence determination. *International Journal of Computer Vision*, 59 (3):259–284, 2004.

[5] O. Enqvist, K. Josephson, and F. Kahl. Optimal correspondences from pairwise constraints. In *International Conference on Computer Vision, ICCV*, pages 1295–1302, 2009.

[6] M.A Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[7] S. Foix, G. Alenya, J. Andrade-Cetto, and C. Torras. Object modeling using a tof camera under an uncertainty reduction approach. In *IEEE International Conference on Robotics and Automation, ICRA*, pages 1306–1312, 2010.

[8] R. Hartley and F. Schaffalitzky. L_∞ Minimization in Geometric Reconstruction Problems. In *Conference on Computer Vision and Pattern Recognition, CVPR*, pages 504–509, 2004.

[9] F. Kahl and R. Hartley. Multiple-view geometry under the L_∞-norm. *IEEE Transactions on Pattern Analylis and Machine Intelligence*, 30(9):1603–1617, 2008.

[10] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[11] F Moreno-Noguer. Deformation and illumination invariant feature point descriptor. In *Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1593–1600, 2011.

[12] F. Moreno-Noguer, V. Lepetit, and P. Fua. Pose priors for simultaneously solving alignment and correspondence. In *European Conference on Computer Vision, ECCV*, volume 5303 of *Lecture Notes in Computer Science*, pages 405–418. Springer, 2008.

[13] K. Ni, H. Jin, and F. Dellaert. Groupsac: Efficient consensus in the presence of groupings. In *International Conference on Computer Vision, ICCV*, pages 2193–2200, 2009.

[14] D. Nistér. Preemptive RANSAC for Live Structure and Motion Estimation. In *International Conference on Computer Vision, ICCV*, pages 199–206, 2003.

[15] A. Penate-Sanchez, J. Andrade-Cetto, and F. Moreno-Noguer. Exhaustive linearization for robust camera pose and focal length estimation. *IEEE Transactions on Pattern Analylis and Machine Intelligence*, 2013.

[16] R. Raguram, J.M. Frahm, and M. Pollefeys. A Comparative Analysis of RANSAC Techniques Leading to Adaptive Real-Time Random Sample Consensus. In *European Conference on Computer Vision, ECCV*, pages 500–513, 2008.

[17] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection.* John Wiley & Sons, Inc., New York, NY, USA, 1987. ISBN 0-471-85233-3.

[18] J. Sanchez, J. Ostlund, P. Fua, and F. Moreno-Noguer. Simultaneous pose, correspondence and non-rigid shape. In *Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1189–1196, 2010.

[19] E. Serradell, M. Özuysal, V. Lepetit, P. Fua, and F. Moreno-Noguer. Combining geometric and appearance priors for robust homography estimation. In *European Conference on Computer Vision, ECCV*, volume 6313 of *Lecture Notes in Computer Science*, pages 58–72. Springer, 2010.

[20] E. Serradell, A. Romero, R. Leta, C. Gatta, and F. Moreno-Noguer. Simultaneous Correspondence and Non-Rigid 3D Reconstruction of the Coronary Tree from Single X-Ray Images. In *International Conference on Computer Vision, ICCV*, pages 850–857, 2011.

[21] J. Shotton, A.Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1297–1304, 2011.

[22] K. Sim and R. Hartley. Removing Outliers Using the L_∞ Norm. In *Conference on Computer Vision and Pattern Recognition, CVPR*, pages 485–494, 2006.

[23] B. Tordoff and D. W. Murray. Guided-MLESAC: Faster Image Transform Estimation by Using Matching Priors. *IEEE Transactions on Pattern Analylis and Machine Intelligence*, 27(10):1523–1535, 2005.

[24] E. Trulls, I. Kokkinos, A. Sanfeliu, and F. Moreno-Noguer. Dense segmentation-aware descriptors. In *Conference on Computer Vision and Pattern Recognition, CVPR*, June 2013.