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

Adrian Penate-Sanchez, Eduard Serradell,
Juan Andrade-Cetto and Francesc Moreno-Noguer

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

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 fixed time. If the percentage of outliers increases, the number of RANSAC loops needed to obtain a correct set of minimal points grows exponentially. 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 [1] to deal with uncalibrated cameras and noisy 3D information.

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. Let us denote by $\mathbf{p}$ a 6-dimensional vector, parameterizing the pose, and let f be the unknown camera focal length. As the 3D-to-2D correspondences are unknown, they need to be retrieved together with the pose and focal length parameters.

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

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

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

is a function that penalizes points whose reprojection error is above a *Max_distance_inlier* threshold to avoid local minima.

We build a hyper-box in the 6-dimensional pose space using the pose priors, 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 $\Sigma_k^{\mathbf{p}}$. 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, $l = 1,\ldots,N_f$.

Our approach lets us handle uncertain 3D models, such as those obtained from a Kinect camera Fig. 1. Each 3D model point $\mathbf{x}_i$ is assigned a covariance $\Sigma_i^{\mathbf{x}}$, computed considering the depth variations of their neighboring points. Those points with larger uncertainties will have less weight in the computation of the solution. We also assign an uncertainty $\Sigma_j^{\mathbf{u}}$ to each 2D measurement. 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 $\{\mathbf{p}_k,\Sigma_k^{\mathbf{p}}\};\{f_l,\Sigma_l^f\}$. 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.

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.

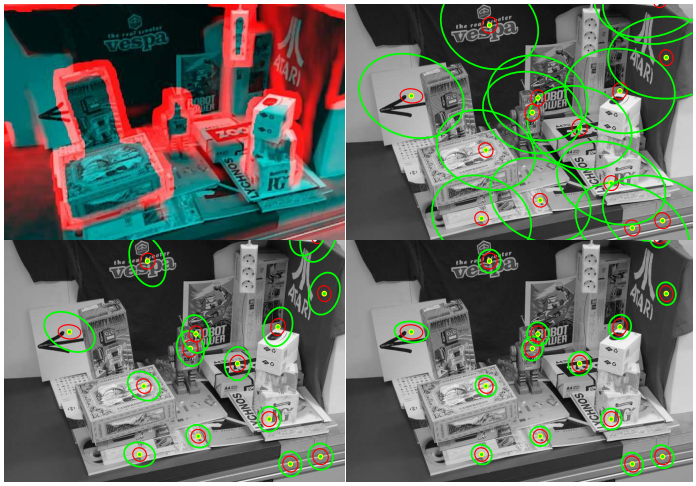

Figure 1: **Up Left:** We detect the uncertain regions –shown in red– computing depth covariances within local neighborhoods. We define a search region to limit the number of candidates **Up Right:** Search region obtained after first projection of the three terms of Eq. 2 independently. **Bottom:** Refinement of the search space, after establishing correspondences.

Once we have defined the set of potential 2D candidates for all the 3D points, we start a hypothesize and test strategy. After hypotheses have 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 Fig. 1. We initialize this step choosing the least ambiguous point

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

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:

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

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.

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.

[1] F. Moreno-Noguer, V. Lepetit, and P. Fua. Pose priors for simultaneously solving alignment and correspondence. In *ECCV*, 2008.