On Sampling Focal Length Values to Solve the Absolute Pose Problem

Torsten Sattler[1], Chris Sweeney[2,⋆], and Marc Pollefeys[1]

[1] Department of Computer Science, ETH Zürich, Zürich, Switzerland
[2] University of California Santa Barbara, Santa Barbara, USA

Abstract. Estimating the absolute pose of a camera relative to a 3D representation of a scene is a fundamental step in many geometric Computer Vision applications. When the camera is calibrated, the pose can be computed very efficiently. If the calibration is unknown, the problem becomes much harder, resulting in slower solvers or solvers requiring more samples and thus significantly longer run-times for RANSAC. In this paper, we challenge the notion that using minimal solvers is always optimal and propose to compute the pose for a camera with unknown focal length by randomly sampling a focal length value and using an efficient pose solver for the now calibrated camera. Our main contribution is a novel sampling scheme that enables us to guide the sampling process towards promising focal length values and avoids considering all possible values once a good pose is found. The resulting RANSAC variant is significantly faster than current state-of-the-art pose solvers, especially for low inlier ratios, while achieving a similar or better pose accuracy.

Keywords: RANSAC, n-point-pose (PnP), camera pose estimation.

1 Introduction

Estimating the absolute camera pose from a set of 2D-3D correspondences, also known as the n-point pose (PnP) problem, is an important step in many Computer Vision applications such as Structure-from-Motion (SfM) [23,25] and image-based localization [12,18,19,21]. Especially for SfM, photo-community collections such as Flickr or Panoramio represent a vast and easily accessible source of data and truly enable large-scale 3D reconstructions [9]. Unfortunately, the EXIF data required to obtain the intrinsic camera calibration of the images is often missing for images obtained from photo sharing websites or is incorrect due to image editing operations applied before uploading the photos [3]. Thus, it is important to estimate both the camera pose and its internal calibration. For the latter, it is often sufficient to estimate only the focal length [2,24].

Computing the camera pose for a calibrated camera is a well-understood problem that has been studied extensively [8,10,14,17]. Given three correspondences between features in an image and points in the 3D model, the camera pose relative to the model can be computed very efficiently by solving a fourth degree

⋆ The first and second author contributed equally to this work.

D. Fleet et al. (Eds.): ECCV 2014, Part IV, LNCS 8692, pp. 828–843, 2014.

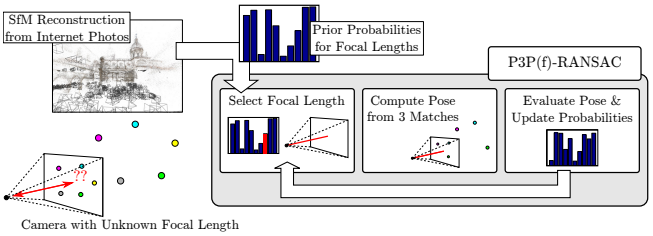

Fig. 1. Illustration of the pose estimation strategy proposed in this paper

polynomial [8,14], resulting in 3-point pose (P3P) solvers that require only about $2\mu s$ on a modern computer [14]. However, estimating the focal length together with the pose is a significantly harder problem. While special configurations such as planar scenes can be handled efficiently [1], computing both quantities generally requires solving a system of multivariate polynomials obtained from four or more 2D-3D correspondences [2,24]. The bottleneck of such approaches is usually the Eigenvalue decomposition of the so-called action matrix and the resulting pose solvers require $46\mu s$ or more for a single instance [4]. Consequently, using such methods inside a RANSAC-loop [8] results in prohibitively long run-times for all but high inlier ratios. In practice, it is thus common to employ pose solvers that achieve similar run-times as P3P [14] but require five or more 2D-3D correspondences [11,16]. As the number of RANSAC iterations grows with both the percentage of false matches and the number of matches required to compute a pose, using such approaches results in significantly longer run-times for low inlier ratios compared to pose solvers using only three or four matches.

In this paper, we consider the problem of estimating the camera pose for a camera with an unknown focal length. Inspired by the brute-force approach of Irschara *et al.* [12], we propose to estimate the focal length by sampling from a discrete set of possible values, followed by computing the pose using the selected focal length instead of simultaneously estimating both quantities. As our main contribution, we propose a novel RANSAC variant, called P3P(f)-RANSAC, that in each iteration randomly selects the focal length value based on the probability of finding a better model for it (*c.f.* Fig. 1). In contrast to [12], which iteratively tests all possible focal length values, we re-estimate the probabilities of each possible focal length value after each RANSAC step using a recursive Bayesian filter. This enables our algorithm to quickly converge toward the focal length closest to the correct value. Consequently, our approach does not necessarily need to evaluate all focal length values, resulting in an average speed-up of more than one order of magnitude compared to [12]. We observe a distribution of focal lengths from photos obtained from photo-sharing websites that allow us to estimate the prior probabilities of the different focal length values, enabling our approach to use importance sampling to find a good pose more quickly. Through experiments on both large-scale SfM datasets and image-based localization tasks, we show that our proposed approach is significantly faster than the state-of-the-art minimal solver [2] while achieving a similar pose accuracy. At the same time,

P3P(f)-RANSAC is faster than a recently published non-minimal solver [16] for low inlier ratios while achieving a higher localization accuracy[1].

The rest of the paper is structured as follows. Sec. 2 reviews related work and Sec. 3 discusses the problem solved in this paper in more detail. We present our novel RANSAC variant combining probabilistic focal length sampling and pose estimation in Sec. 4. Sec. 5 then evaluates the resulting approach.

2 Related Work

Estimating the camera pose from n 2D-3D matches is commonly known as the n-point-pose (PnP) problem and algorithms solving this problem are consequently called pose solvers. In case that the camera is calibrated, three correspondences are sufficient to estimate the pose and P3P solvers usually proceed by first estimating the position of the three points in the local coordinate system of the camera before estimating the transformation from the global into the local system from these positions [10]. Recently, Kneip *et al.* proposed a method that directly estimates the camera pose in the global coordinate frame [14]. Similar to [8], their method needs to solve a 4th degree univariate polynomial, which can be done in closed form, resulting in run-times of around $2\mu s$. If the gravity direction is known, the pose estimation problem can be simplified such that only two matches are required [15]. While these pose solvers are used inside a RANSAC-loop to robustly handle outliers, it is common to afterwards use the inlier matches to refine the pose through a general PnP algorithm [17].

In the case that the camera calibration is unknown, the classic 6-point direct linear transform algorithm estimates both the full internal and the external calibration of the camera from six 2D-3D matches by computing the SVD of a 12×12 matrix [11]. Triggs generalized this approach to incorporate prior knowledge about some calibration parameters, resulting in 4-point and 5-point solvers [24]. Similar to the 6-point solver, they cannot handle planar point configurations. Handling general configuration usually results in system of multivariate polynomials [2,3,5,13,24]. Bujnak *et al.* proposed such an approach for the case that only the focal length is unknown [2]. Using four 2D-3D matches, their method needs to perform Gauss-Jordan elimination on a 154×180 matrix followed by computing the Eigenvalues of a 10×10 action matrix, resulting in run-times of $100\mu s$ or more. A faster solver can be obtained using an automatically generated elimination template together with a more efficient way to compute the Eigenvalues, reducing the run-time to $46\mu s$ [4]. [13] show that four correspondences are enough to estimate both the focal length and a radial distortion parameter for general point configurations. However, handling planar and non-planar scenes seperately results in significantly faster run-times [3]. While such minimal solvers still require about $260\mu s$ or more, Kukelova *et al.* recently proposed a non-minimal 5-point solver that only relies on linear algebra and is thus orders of magnitude faster while still recovering the focal length and up to three radial distortion parameters [16].

[1] We make our source code available at `http://people.inf.ethz.ch/sattlert`

Similar to the approach proposed in this paper, Irschara *et al.* [12] repeatedly apply RANSAC with a P3P solver to each focal length in a set of focal length values to obtain the pose for an uncalibrated camera rather than estimating the focal length directly. The focal length value that produces the best pose is then chosen as the focal length for the camera. However, we show that our probabilistic formulation is much more efficient than the brute-force method proposed by [12]. The key idea of our RANSAC variant is to randomly sample the focal length in each iteration according to a given probability distribution. [22] use a similar RANSAC algorithm to calibrate a network of cameras from silhouettes extracted from video. In each iteration, they randomly select two directions in two images to obtain a hypothesis for the epipoles, which is used to recover the full fundamental matrix. This enables them to recover the epipolar geometry even though they cannot establish reliable point correspondences between the silhouettes detected in different images. While [22] sample according to a fixed distribution, we re-estimate the probabilities after each RANSAC iteration to incorporate information from previous rounds.

3 Problem Formulation

In this paper, we want to solve the problem of estimating the pose for a camera with an unknown focal length from a given set $\mathcal{M} = \{(\mathbf{x}, \mathbf{X}) \mid \mathbf{x} \in \mathbb{R}^2, \mathbf{X} \in \mathbb{R}^3\}$ of 2D-3D matches. Assuming that the principal point coincides with the center of the image, we are thus trying to determine the focal length $f \in \mathbb{R}$ and the rotation $\mathtt{R} \in \mathbb{R}^{3\times3}$ and translation $\mathbf{t} \in \mathbb{R}^3$ such that

$$\alpha \cdot \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} [\mathtt{R}|\mathbf{t}] \cdot \begin{pmatrix} \mathbf{X} \\ 1 \end{pmatrix} \quad \text{for some scalar } \alpha > 0 \tag{1}$$

holds for all matches $(\mathbf{x}, \mathbf{X}) \in \mathcal{M}$, *i.e.*, that each 3D point $\mathbf{X}$ is projected onto its corresponding image position $\mathbf{x}$. In practice, some of the matches will be wrong due to imperfections in the matching process. The most common strategy to robustly handle wrong matches is to apply a PnP solver that computes the pose from n matches inside a RANSAC-loop [8]. RANSAC iteratively selects a random subset of size n from the given matches and uses it to estimate the camera pose. The pose is then evaluated on all matches, where a match is considered as an *inlier* to the pose if the reprojection error is below a given threshold and as an *outlier* otherwise. The model with the highest number of inliers is considered as the current best estimate of the correct camera pose. RANSAC terminates once the probability of having missed the correct pose falls below the desired failure probability η. Assuming that each all-inlier sample allows us to estimate the correct pose, this probability may be expressed as

$$(1 - \varepsilon^{*n})^k < \eta \ , \tag{2}$$

where k is the number of samples generated so far and ε^* is the *inlier ratio*, *i.e.*, the ratio of inliers among all matches, for the current best model. Thus, the maximal number of iterations required for a given inlier ratio ε is

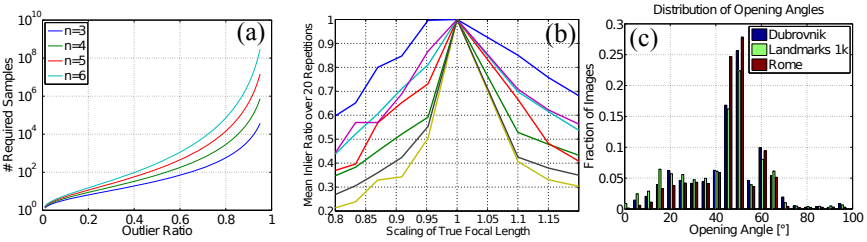

Fig. 2. (a) The number of RANSAC iterations required to ensure that the correct model is found with 99% probability for different PnP solvers. (b) The focal length accuracy required to recover most of the inliers strongly varies between different cameras. Yet, the inlier ratio decreases monotonically on both sides of the optimal focal length value. (c) Histograms of opening angles from images in the Dubrovnik [18], Landmarks 1k [19], and Rome [18] datasets.

$$k_{\max} = \log \eta / \log \left(1 - \varepsilon^n\right) \ . \tag{3}$$

The probability of selecting an all-inlier sample is maximized by minimizing n. However, the minimal 4-point solver (P4Pf) [4] for the problem of estimating both the pose and the focal length requires $46\mu s$, which is prohibitively expensive for low inlier ratios where many RANSAC iterations are required. Faster pose solvers such as the P5Pfr method [16] that estimates the pose, focal length, and radial distortion of the camera from five matches exist. However, using a non-minimal n reduces the probability of selecting an all-inlier sample exponentially, resulting in a significant increase in the number of required iterations for low inlier ratios (*c.f.* Fig. 2(a)). Instead of using a non-minimal solver, we propose to use a 3-point solver that estimates the pose for a given focal length f [14] and select f from a pre-defined set $\mathcal{F}$ of focal length values. This strategy offers the possible advantage of requiring fewer iterations than RANSAC with P4Pf (*c.f.* Fig. 2(a)) and faster pose computation times by using the P3P solver.

Evaluating all focal length values in $\mathcal{F}$ independently from each other as proposed by [12] will require at least $|\mathcal{F}| \cdot k_{\max}(f_{\text{gt}})$ iterations in total, where $k_{\max}(f_{\text{gt}})$ is the maximum number of iterations required to confidently compute the pose when using the ground truth focal length. Consequently, the approach from [12] will only be more efficient than using RANSAC with P4Pf or P5Pfr if $|\mathcal{F}|$ is smaller than the difference in the pose solver time or the difference in the number of required iterations, respectively. Notice that using quantized focal length values will invariably result in a lower pose accuracy. Regardless, as long as we are able to recover most of the inliers we will be able to obtain a better pose by applying P4Pf on the resulting inliers with only a small run-time overhead as very few sampling steps will be needed. Unfortunately, the sampling density required to guarantee that we can select a focal length value close enough to f_{gt} to recover most of the inliers strongly depends on the depth-variation of the scene observed by the camera. This can be seen in Fig. 2(b), as we observe

Algorithm 1 P3P(f)-RANSAC

Given: Set $\mathcal{M}$ of 2D-3D matches, desired failure probability η, **set $\mathcal{F}$ of focal length values with prior probabilities $P_{\mathbf{prior}}(f)$ for all $f \in \mathcal{F}$**
initialize sampling probability $P_{\mathbf{sample}}(f) = P_{\mathbf{prior}}(f)$ for all $f \in \mathcal{F}$
while probability of having missed the correct pose $\geq \eta$ **do**
  randomly select focal length $f \in \mathcal{F}$ according to $P_{\mathbf{sample}}$
  draw random sample $s \subset M$ of size 3
  estimate pose $[\mathtt{R}|\mathbf{t}]$ from s with a P3P solver **using** f
  evaluate pose hypothesis $\theta = (f, [\mathtt{R}|\mathbf{t}])$ on $\mathcal{M}$
  if new best model found **then**
    $\theta^* = (f, [\mathtt{R}|\mathbf{t}])$
  Update probabilities $P_{\mathbf{sample}}$
  Re-estimate probability of having missed the correct pose
Return: θ^*

different sensitivities on the focal length accuracy for different cameras. Thus, we need a rather dense sampling in order to handle all types of scenes, resulting in a large set $\mathcal{F}$. In order to maintain fast run-times when using a large set of values, we model the dependencies between the different focal lengths, enabling us to avoid evaluating all focal length values for at least $k_{\max}(\varepsilon_{\text{gt}})$ steps. This can be done by exploiting a *key observation* that can be made from Fig. 2(b): The maximal inlier ratio obtained by RANSAC for each focal length value decreases monotonically with the distance to f_{gt}. Given the focal length used to generate the current best pose with the highest inlier count, f^*, this observation allows us to model the probability of finding a pose with a higher inlier ratio using another focal length f as a function of $|f - f^*|$.

4 Interdependent Probabilistic Focal Length Sampling

The main idea of our novel pose estimation approach is to use focal length sampling and a P3P solver [14] in order to estimate a hypothesis for the camera pose from $n=3$ 2D-3D correspondences instead of computing the pose and focal length simultaneously from four matches or more. Once we have found a good pose with a high inlier ratio for a focal length f^*, it becomes very unlikely that focal length values f far away from f^* can be used to estimate a better pose (*c.f.* Fig 2(b)). The central idea behind our approach is thus to preferably select focal length values that have a high likelihood of yielding a pose with a larger number of inliers than the current best estimate. This naturally leads to a probabilistic formulation of the problem of selecting good focal length values. This probabilistic formulation in turn enables us to exploit the fact that certain focal length values are much more likely to be correct than others. Alg. 1 outlines the resulting RANSAC variant, where differences to the classical RANSAC algorithm [8] are highlighted. Besides the 2D-3D matches and the failure probability η, our approach requires a set $\mathcal{F}$ of focal length values with associated prior probabilities as an additional input. These priors are then used to initialize

the probability distribution that we use for selecting the focal length value f in Line 3 of Alg. 1. After using P3P to generate a pose hypothesis from f and three randomly selected matches, the hypothesis is evaluated on all matches and the current best pose estimate is updated if necessary. Finally, we use a recursive Bayesian filter to re-estimate the probability distribution used for sampling the focal length to reflect the fact that the current iteration might influence the likelihood of finding a better pose for all other focal length values.

In the following, we will refer to our algorithm as P3P(f)-RANSAC, as it uses a P3P solver inside of a RANSAC loop, where the focal length value f is obtained via parameter sampling. Similarly, we will refer to RANSAC-loops using any other PnP solver as PnP-RANSAC.

In Sec. 4.1, we briefly explain how to obtain the prior probabilities for the focal length values from $\mathcal{F}$. As the main contribution of this paper, Sec. 4.2 derives the probability distribution used for sampling the focal length values and our strategy for re-estimating the sampling probabilities. Finally, Sec. 4.3 argues that using early model rejection techniques [6,7] is crucial for our RANSAC variant in order to offer faster run-times than P4Pf and P5Pfr.

4.1 Obtaining the Prior Probabilities

The focal length of a camera mainly depends on the type of camera and the zoom-level used to take the picture. In this paper, we consider pose estimation scenarios in which a large variety of camera types is used, as is the case in large-scale SfM reconstructions from images downloaded from Flickr [23,9]. Since some camera types are much more popular than others[2], not all focal length values are equally likely to occur. The cameras contained in a large-scale SfM reconstruction of community collection photos thus give us an approximation to the probability distribution of focal length values. However, notice that obtaining prior probabilities for focal length values is an ill-posed problem as the focal length depends on the image resolution. In contrast, the maximal opening angle $\alpha_{\max}$ of a camera with focal length f, width w, and height h, related by

$$\tan\left(\alpha_{\max}/2\right) = \frac{\max\left(w,h\right)}{2 \cdot f} \quad , \tag{4}$$

is independent of the image resolution. Thus, we predetermine a set of opening angle values from cameras contained in large-scale SfM reconstructions of unordered image collections [18,19]. We transform the opening angles to focal length values via Eqn. 4 (based on the resolution of the image being localized) before applying P3P(f)-RANSAC. Fig. 2(c) shows the distribution of opening angles for three such datasets, Dubronik (6k images) [18], Rome (15k images) [18], and the Landmarks 1k dataset (205k images) [19]. The distribution of opening angles is consistent across all datasets, indicating that the resulting distributions are a good representation of images taken in the real world. Still, we will show in Sec. 5.2 that the choice of priors is not a crucial parameter.

[2] `https://www.flickr.com/cameras`

4.2 Obtaining and Re-estimating the Sampling Probabilities

Ideally, the probability $P_{\text{sample}}(f)$ of selecting a focal length f should be proportional to the likelihood of obtaining a pose estimate with an inlier ratio $\varepsilon(f)$ that is larger than the inlier ratio ε^* of the current best pose estimate θ^* obtained for focal length f^*. Consequently, we model the sampling probability as

$$P_{\text{sampling}}(f) = \frac{P(\varepsilon(f) > \varepsilon^* \mid f) \cdot P_{\text{prior}}(f)}{\sum_{f' \in \mathcal{F}} P(\varepsilon(f') > \varepsilon^* \mid f') \cdot P_{\text{prior}}(f')} \ , \tag{5}$$

where $P(\varepsilon(f) > \varepsilon^* \mid f)$ is the probability of finding a better model using the focal length f. As is common in practice, we assume that we can obtain an inlier ratio of at least ε_0 in order to limit the maximal number of RANSAC iterations, *i.e.*, we assume $\varepsilon^* = \varepsilon_0$ until we find a pose with an inlier ratio $> \varepsilon_0$.

In the following, we first derive $P(\varepsilon(f) > \varepsilon_0 \mid f)$ for the case that all models found so far have an inlier ratio of at most ε_0. In this case, we have not yet found a good model and thus have to treat all focal length values independently. We then show that the case of having found a good model with $\varepsilon^* > \varepsilon_0$, in which case $P(\varepsilon(f) > \varepsilon^* \mid f)$ depends on the current best pose θ^*, seamlessly integrates into our definition of the probabilities.

Case 1: $\varepsilon^* = \varepsilon_0$**.** Using the termination criterion from Eqn. 2, we express the maximal inlier ratio $\varepsilon_{\max}(f)$ that we have missed with probability $\geq \eta$ in terms of the number of random samples $k(f)$ generated so far for focal length f:

$$\varepsilon_{\max}(f) = \sqrt[3]{1 - \sqrt[k(f)]{\eta}} \ . \tag{6}$$

Since we are only required to compute the correct pose with probability $\geq \eta$, the probability $P(\varepsilon(f) > \varepsilon_0 \mid f)$ of finding a model with a higher inlier ratio is directly related to the probability that the number of correct matches in $\mathcal{M}$ is in the range $(\varepsilon_0 \cdot |\mathcal{M}|, \varepsilon_{\max}(f) \cdot |\mathcal{M}|]$. Notice that the probability of finding a wrong match only depends on the matching algorithm and the structure of the 3D model [21], and *not* on the pose estimation strategy itself. Since this probability can be estimated empirically from training data, we can assume without loss of generality that we know the cumulative distribution function $\text{cdf}(\varepsilon)$ over the inlier ratios for the given matching algorithm and 3D model. Thus, we can express the probability of finding a better model for f as

$$P(\varepsilon(f) > \varepsilon_0 \mid f) = \text{cdf}(\max(\varepsilon_{\max}(f), \varepsilon_0)) - \text{cdf}(\varepsilon_0) \ . \tag{7}$$

Under the reasonable assumption that $\text{cdf}(\varepsilon)$ is strictly increasing, *i.e.*, that all inlier ratios occur with a non-zero probability, we have $P(\varepsilon(f) > \varepsilon_0 \mid f) = 0$ only if $\varepsilon_{\max}(f)) \leq \varepsilon_0$. Consequently, P3P(f)-RANSAC will terminate after $|\mathcal{F}| \cdot k_{\max}(\varepsilon_0)$ iterations, *i.e.*, if no pose with inlier ratio greater than ε_0 can be found with a probability of at least η.

Case 2: $\varepsilon^* > \varepsilon_0$**.** Note that $P(\varepsilon(f) > \varepsilon^* \mid f)$ not only depends on the inlier ratio ε^* but also on the value of the focal length f^* used to compute the current

best hypothesis θ^*. If f^* is close to the correct focal length f_{gt}, then focal length values far away from f^* are much less likely to result in better pose hypotheses than values close to f^*. This behavior can also be observed in Fig. 2(b), which shows that the inlier ratio decreases monotonically with the distance to the correct focal length when applying RANSAC on correct matches only. While outlier matches might cause local maxima, we found that this relation is still a very good model in practice. Since a similar behavior has been observed for other estimation problems [20], we thus use the following simplifying assumption to derive the sampling probabilities.

Assumption 1. *Let $\varepsilon(f)$ be the maximal inlier ratio that can be obtained for focal length f and let f_{gt} be the correct focal length. For focal length values f and f' with $|f_{gt} - f'| < |f_{gt} - f|$, $\varepsilon(f) \leq \varepsilon(f') \leq \varepsilon(f_{gt})$ should hold.*

Without loss of generality, consider the focal length $f < f^*$. If f is closer to f_{gt} than f^*, Assumption 1 implies that we should be able to find an inlier ratio of at least ε^* for all $f' \in \mathcal{F} \cap [f, f^*)$. Let $\mathcal{F}(f, f^*) = \mathcal{F} \cap [f, f^*)$ be the set of corresponding focal length values and let $P(\varepsilon(\mathcal{F}(f, f^*)) > \varepsilon^* \mid f)$ denote the probability of finding a better pose in the range $[f, f^*)$, then we have

$$P(\varepsilon(f) > \varepsilon^* \mid f) \quad \leq \quad P(\varepsilon(\mathcal{F}(f, f^*)) > \varepsilon^* \mid f) \ . \tag{8}$$

The maximal inlier ratio in this range of focal lengths that we have missed with a probability of at least η is again given by

$$\varepsilon_{\max}(\mathcal{F}(f, f^*)) = \sqrt[3]{1 - \sqrt[k(\mathcal{F}(f,f^*))]{\eta}} \ , \tag{9}$$

where $k(\mathcal{F}(f, f^*)) = \sum_{f' \in \mathcal{F}(f,f^*)} k(f')$ is the sum over all samples generated for the focal lengths from the considered range. As in Case 1, we thus obtain

$$P(\varepsilon(f) > \varepsilon^* \mid f) = \text{cdf}(\max(\varepsilon_{\max}(\mathcal{F}(f, f^*)), \varepsilon^*)) - \text{cdf}(\varepsilon^*) \ . \tag{10}$$

This predict-and-update strategy is a recursive Bayesian filter. Note that we again have $P(\varepsilon(f) > \varepsilon^* \mid f) = 0$ only if the probability of finding a better pose for f drops above the confidence threshold η, *i.e.*, P3P(f)-RANSAC essentially uses the same termination criterion as original RANSAC, offering the same guarantees on the quality of the pose.

Behavior of the Proposed Sampling Strategy. As long as no pose with an inlier ratio above ε_0 is found (Case 1), P3P(f)-RANSAC essentially uses importance sampling to select promising focal length values. As soon as a good model with inlier ratio above ε_0 is found (Case 2), P3P(f)-RANSAC is able to model the dependencies between focal length values, allowing it to quickly focus on a smaller range of focal length values that are most likely to be correct. This behavior is illustrated in Fig. 3. At the same time, our sampling strategy is able to escape local maxima since all focal length values that could lead to a better pose have a non-zero probability of being selected.

Implementation Details. Each focal length value is used for at most $k_{\max}(\varepsilon_0)$ samples. Since both Eqn. 6 and Eqn. 9 only depend on the number of iterations

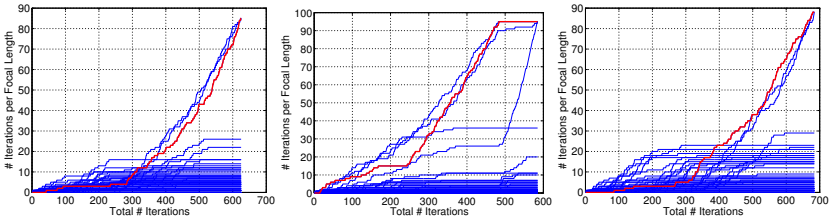

Fig. 3. The number of iterations in which each of the 100 focal length values is selected, plotted over the iterations of P3P(f)-RANSAC for three cameras from the Dubrovnik dataset and an outlier ratio of 50%. The focal length value closest to the true focal length of each camera is highlighted in red. As can be seen, P3P(f)-RANSAC is able to quickly identify a subset of promising focal lengths while ignoring all other values.

and not on ε^*, we can use a lookup table to determine the maximal inlier ratio. We represent the (empirically determined) cumulative distribution function $\text{cdf}(\varepsilon)$ as a discrete set of values. For any inlier ratio ε', we use linear interpolation to compute $\text{cdf}(\varepsilon')$ to guarantee that our discrete representation is still strictly increasing, which prevents P3P(f)-RANSAC from terminating too early.

4.3 Integrating Early Model Rejection

The P3P solver can compute the pose from three 2D-3D matches in $2\mu s$ [14] while the fastest P4Pf solver takes $46\mu s$ [4]. Consequently, P3P(f)-RANSAC should be able to perform 23 times more sampling steps while still being faster than P4Pf-RANSAC. However, evaluating the computed pose on the set of matches also has a significant impact on the run-time of a single RANSAC iteration. Since evaluating a pose takes around $20-50\mu s$ (or more for images with a large number of matches), P3P(f)-RANSAC can be at most $2-3$ times faster than P4Pf-RANSAC when evaluating each pose on all matches. Obviously, we do not need to fully evaluate poses generated from non-all-inlier samples or with a wrong focal length value. We can thus use approaches that terminate the pose evaluation once it becomes likely that the current pose will not have an inlier ratio higher than ε^* [6,7]. We chose to use the simple $T_{d,d}$ test, which evaluates a pose on all matches only if d randomly selected matches are inlier to the pose, with $d=1$ as proposed in [6]. As a result of applying this $T_{1,1}$ test, we need to draw $n=4$ matches in each iteration of P3P(f)-RANSAC, increasing the number of required iterations (*c.f.* Eqn. 3). At the same time, it becomes rather unlikely that any pose estimated from a focal length far away from the correct value, even if it was estimated only from correct matches, is evaluated on all correspondences since significantly fewer correct matches are inliers to such poses (*c.f.* Fig. 2(b)). As a consequence, only a small fraction of all generated poses need to be fully estimated, resulting in a significant speed-up.

5 Experimental Evaluation

In the following, we evaluate the performance of our proposed method both on synthetic and real-world data. For all experiments, we use the Landmarks 1k dataset [19], reconstructed from 205k Flickr images, to learn the probability distribution for 100 equally spaced opening angles, which we then transform into focal length values for any image with a given width and height.

Using realistic focal lengths is an important part of our experiments, since our algorithm utilizes the distribution of likely focal lengths to inform our RANSAC scheme. In order to obtain realistic focal length values, and realistic 2D-3D matches, for our synthetic experiments, we use two large-scale SfM reconstructions and generate pixel-perfect 2D-3D correspondences by reprojecting the 3D points into the images in which they were observed. The Rome model [18] consists of 15k database images and ∼4M points, while ∼1.9M points were reconstructed from 6k images to create the Dubrovnik model [18]. The scale for the latter model is known, allowing us to measure the localization accuracy on the Dubrovnik dataset in meters. Both datasets form a standard benchmark for image-based localization tasks [18,19,21] and we thus evaluate the performance on real-world data of our approach in this application scenario. For both datasets we use a cdf learned from inlier ratios observed on the Dubrovnik dataset.

For our experiments, we used the publicly available implementations of P3P [14] and P4Pf [2] and our own implementation of the P5Pfr solver [16].

5.1 Experiments with Synthetic Data

We conducted two synthetic experiments to measure the performance of our algorithm under increased levels of image noise and outlier ratios.

Image Noise. We measured our algorithm's robustness against image noise by adding increasing levels of Gaussian pixel noise to the 2D positions of the perfect 2D-3D correspondences obtained by reprojecting the 3D points. We tested image noise levels of 0, 0.1, 0.5, 1.0, and 2.0 pixels. Fig. 4 compares the performance of our approach with P4Pf-RANSAC. For all levels of image noise, P4Pf achieves slightly lower rotation, translation, and focal length errors, though the errors are comparable. This indicates that our algorithm is able to estimate the pose and focal length with high precision and is thus robust to noise, which is important for real-world data.

Outlier Ratio. The key idea of our approach is to use the faster P3P solver to estimate camera poses more efficiently while avoiding a brute-force search through all possible focal length values through our novel sampling scheme. In this experiment, we evaluate the robustness of our approach to high outlier ratios. We again use the perfect matches from the Dubrovnik dataset, with 1 pixel of Gaussian noise added to the reprojected points, and create outliers by adding new image points with correspondences to 3D points that were not observed in the image until the desired outlier ratio is achieved.

Fig. 5 shows the performance of our P3P(f) approach and P4Pf-RANSAC for increasing levels of outlier ratios. We plot the median position errors, inlier

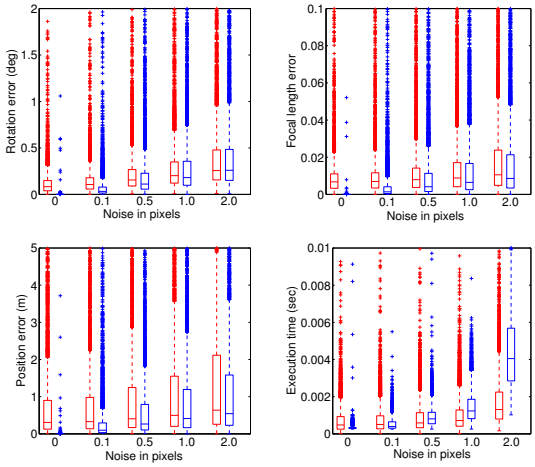

Fig. 4. Performance of our algorithm (red) and P4Pf [2] (blue) are compared for increased levels of image noise. Our algorithm has comparable performance to P4Pf for rotation, position, and focal length errors for all levels of noise. Despite requiring more iterations, our algorithm has a lower run-time than P4Pf as the image noise increases.

ratios, and execution times. As can be seen, our algorithm is able to handle low-inlier scenarios and still produce results that are nearly as accurate as P4Pf while being several orders of magnitude faster. These results demonstrate that Assumption 1 holds well enough even in the presence of outliers. For tasks such as image-based localization, being able to handle low-inlier scenarios accurately and efficiently is extremely important.

5.2 Experiments on Real Data

As a final experiment, we compare the performance of our algorithm to P3P, P4Pf, and P5Pfr in an image-based localization task [18,19,21]. We use two versions of our algorithm: One with focal length priors obtained from the Landmarks 1K dataset, and one with no learned priors (*i.e.* uniform priors). We use the efficient, publicly available localization method of [21] to obtain 2D-3D matches for the 800 and 1000 query images available for the Dubrovnik and Rome datasets, respectively. All query images were obtained by removing cameras from larger SfM reconstructions, providing ground truth positions for the query images. Notice that we do not use perfect correspondences in these experiments.

The results for the Rome dataset are shown in Fig. 6. Algorithms that computed focal length in addition to pose are able to recover noticeably more inliers than the P3P method that was used with ground truth focal lengths values as we did not account for radial distortion. As expected, all of the algorithms are slower than P3P. Our algorithm performed much faster than P4Pf in all cases.

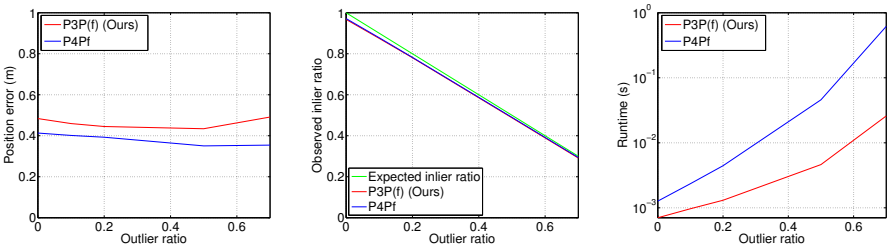

Fig. 5. The median position error, inlier ratio, and run-time was measured while increasing the outlier ratio from 0 to 0.7. Both algorithms are able to recover high quality poses (left) and almost all expected inliers (middle). Our algorithm has a much lower run-time than P4Pf (right) as the outlier ratio increases due to using a faster solver. This is a major advantage of our algorithm in low-inlier scenarios.

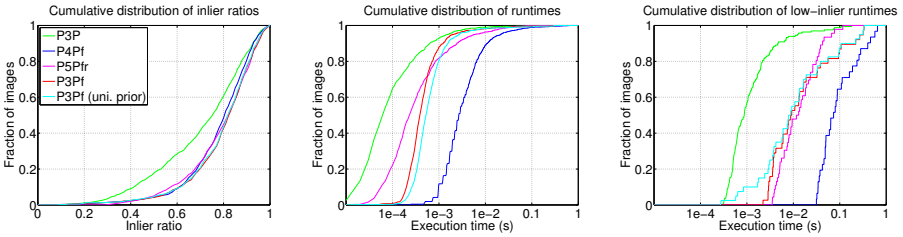

Fig. 6. Localization results from the Rome dataset [18] are shown. Our P3P(f)-RANSAC algorithm is able to recover more inliers than P3P used with ground truth focal lengths from Bundler, and a comparable amount to P4Pf and P5Pfr (left). Our algorithm has an execution time that is nearly one order of magnitude faster than P4Pf (center), despite running for more iterations. In low-inlier cases (inlier ratio ≤ 0.5), our algorithm is significantly faster than alternative algorithms (right).

As shown in Fig. 6, our approach is faster than P5Pfr for most low-inlier cases as it requires fewer matches per sample and thus fewer iterations per focal length.

Tab. 1 shows the position errors of each method on the Dubrovnik dataset, where we can measure distances in meters. The median position error of each camera was recorded over 100 trials for each of the methods. All methods are able to localize almost all images, and our method gives position errors that are comparable to or only slightly higher than P4Pf, which has the lowest errors of all algorithms. P3P(f) achieves better localization accuracy than P5Pfr. As can be also seen in Tab. 1, our method is on average over an order of magnitude faster than P4Pf. At the same time, P3P(f) is consistently faster than P4Pf on all quantiles while being faster than P5Pfr for images with lower inlier ratios. Notice that our P4Pf implementation requires $115\mu s$ compared to the $46\mu s$ required by [4]. Yet, our approach is on average more than 7 times faster than when using the solver from [4] and still achieves faster quantile run-times. On average, P3P(f) is only 1.39 times slower than P3P, even though it requires no knowledge about the focal length, making it well suited for SfM and localization applications.

Table 1. The position errors and localization times measured on the Dubrovnik dataset for an image-based localization task. Besides the results obtained by our approach using the learnt priors for the focal lengths, we also include results for an uniform prior.

Solver	# loc. images	Localization Accuracy [m]					Localization Times [ms]			
		Mean [m]	Quantiles [m]				Mean [ms]	Quantiles [ms]		
			25%	50%	75%	90%		50%	75%	90%
P3P (exact focal)	792	40.3	1.0	7.6	26.4	111.8	**1.21**	**0.20**	**1.00**	3.01
P4Pf	795	38.7	**0.4**	**1.3**	**4.7**	**20.1**	32.09	4.84	10.78	28.73
P5Pfr	**796**	227.2	0.5	2.0	31.3	200.9	6.02	0.54	3.07	16.44
P3P(f) (Ours)	795	**20.8**	**0.4**	1.6	5.4	27.6	1.68	0.68	1.27	**2.72**
P3P(f) uniform prior	795	28.1	0.5	1.7	5.9	24.3	1.89	0.85	1.46	3.08

Tab. 1 and Fig. 6 also show results obtained using a uniform prior on the focal lengths. As can be seen, our method benefits from using a good prior but performs only slightly worse otherwise. This demonstrates that our novel sampling scheme is the main reason for why P3P(f)-RANSAC succeeds.

6 Conclusion

In this paper, we have proposed a novel approach, termed P3P(f)-RANSAC, for efficiently estimating the pose of a camera with unknown focal length inside a RANSAC loop. Instead of computing the focal length using a minimal solver, our approach samples focal length values according to a probability distribution and then uses the significantly faster P3P solver to estimate the pose of the now calibrated camera. As the main contribution, we have proposed a novel sampling scheme that is able to model the probability of finding a pose better than the current best estimate for all focal length values. As a consequence, our approach is able to avoid evaluating all values and focus on the more promising candidates while offering the same guarantees as RANSAC in the presence of outliers. We have shown that our algorithm achieves a similar pose accuracy as previous pose solvers while achieving significantly faster run-times. These results challenge the notion that using minimal solvers is always an optimal strategy. While this paper focusses on the absolute pose problem, we plan to explore the use of our framework for other pose estimation problems in future work.

Acknowledgements. This work was supported in part by NSF Grant IIS-1219261, NSF Graduate Research Fellowship Grant DGE-1144085, the CTI Switzerland grant #13086.1 PFES-ES 4DSites, and the European Union's Seventh Framework Programme (FP6/2007-2013) under grant #269916 (V-Charge).

References

1. Abidi, M.A., Chandra, T.: A New Efficient and Direct Solution for Pose Estimation using Quadrangular Targets: Algorithm and Evaluation. PAMI 17(5), 534–538 (1995)

2. Bujnak, M., Kukelova, Z., Pajdla, T.: A General Solution To The P4P Problem for Camera With Unknown Focal Length. In: CVPR (2008)
3. Bujnak, M., Kukelova, Z., Pajdla, T.: Robust Focal Length Estimation by Voting in Multi-view Scene Reconstruction. In: Zha, H., Taniguchi, R.-i., Maybank, S. (eds.) ACCV 2009, Part I. LNCS, vol. 5994, pp. 13–24. Springer, Heidelberg (2010)
4. Bujnak, M., Kukelova, Z., Pajdla, T.: New efficient solution to the absolute pose problem for camera with unknown focal length and radial distortion. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part I. LNCS, vol. 6492, pp. 11–24. Springer, Heidelberg (2011)
5. Bujnak, M., Kukelova, Z., Pajdla, T.: Making Minimal Solvers Fast. In: CVPR (2012)
6. Chum, O., Matas, J.: Randomized RANSAC with T(d,d) test. In: BMVC (2002)
7. Chum, O., Matas, J.: Optimal Randomized RANSAC. PAMI 30(8), 1472–1482 (2008)
8. Fischler, M., Bolles, R.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Comm. ACM 24(6), 381–395 (1981)
9. Frahm, J.-M., et al.: Building rome on a cloudless day. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 368–381. Springer, Heidelberg (2010)
10. Haralick, R., Lee, C.N., Ottenberg, K., Nölle, M.: Review and analysis of solutions of the three point perspective pose estimation problem. IJCV 13(3), 331–356 (1994)
11. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge Univ. Press (2004)
12. Irschara, A., Zach, C., Frahm, J.M., Bischof, H.: From Structure-from-Motion Point Clouds to Fast Location Recognition. In: CVPR (2009)
13. Josephson, K., Byröd, M.: Pose Estimation with Radial Distortion and Unknown Focal Length. In: CVPR (2009)
14. 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: CVPR (2011)
15. Kukelova, Z., Bujnak, M., Pajdla, T.: Closed-form solutions to the minimal absolute pose problems with known vertical direction. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part II. LNCS, vol. 6493, pp. 216–229. Springer, Heidelberg (2011)
16. Kukelova, Z., Bujnak, M., Pajdla, T.: Real-Time Solution to the Absolute Pose Problem with Unknown Radial Distortion and Focal Length. In: ICCV (2013)
17. Lepetit, V., Moreno-Noguer, F., Fua, P.: EPnP: An Accurate O(n) Solution to the PnP Problem. IJCV 81(2), 155–166 (2009)
18. 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)
19. Li, Y., Snavely, N., Huttenlocher, D., Fua, P.: Worldwide Pose Estimation Using 3D Point Clouds. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 15–29. Springer, Heidelberg (2012)
20. Nister, D.: An Efficient Solution to the Five-Point Relative Pose Problem. PAMI 26(6), 756–770 (2004)

21. Sattler, T., Leibe, B., Kobbelt, L.: Improving Image-Based Localization by Active Correspondence Search. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 752–765. Springer, Heidelberg (2012)
22. Sinha, S.N., Pollefeys, M.: Camera Network Calibration and Synchronization from Silhouettes in Archived Video. IJCV 87(3), 266–283 (2010)
23. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3D. In: SIGGRAPH (2006)
24. Triggs, B.: Camera Pose and Calibration from 4 or 5 Known 3D Points. In: ICCV (1999)
25. Wu, C.: Towards Linear-Time Incremental Structure from Motion. In: 3DV (2013)