Exploiting Uncertainty in Regression Forests for Accurate Camera Relocalization

Julien Valentin[1], Matthias Nießner[2], Jamie Shotton[3], Andrew Fitzgibbon[3], Shahram Izadi[3], Philip Torr[1]
[1]Oxford University [2]Stanford University [3]Microsoft Research

Recent advances in camera relocalization use predictions from a regression forest to guide the camera pose optimization procedure. In these methods, each tree associates one pixel with a point in the scene's 3D world coordinate frame. In previous work, these predictions were point estimates and the subsequent camera pose optimization implicitly assumed an isotropic distribution of these estimates. In this paper, we train a regression forest to predict mixtures of anisotropic 3D Gaussians and show how the predicted uncertainties can be taken into account for continuous pose optimization. Experiments show that our proposed method is able to relocalize up to 40% more frames than the state of the art. The main contributions of this work are (i) the extension of the state of the art on RGB-D camera relocalization by modeling and minimizing uncertainties for regression tree induction and predictions performed by the regression forest; and (ii) leveraging these uncertainties in order to provide for improved relocalization without using explicit models of the scenes.

The proposed camera relocalization approach consists of two major components: (i) a regression forest trained on RGB-D input data to predict anisotropic Gaussian mixtures of 3D scene coordinates; and (ii) a continuous pose optimization leveraging the anisotropic Gaussian mixtures predicted by the forest. An overview of the complete relocalization pipeline is shown in Fig. 1.

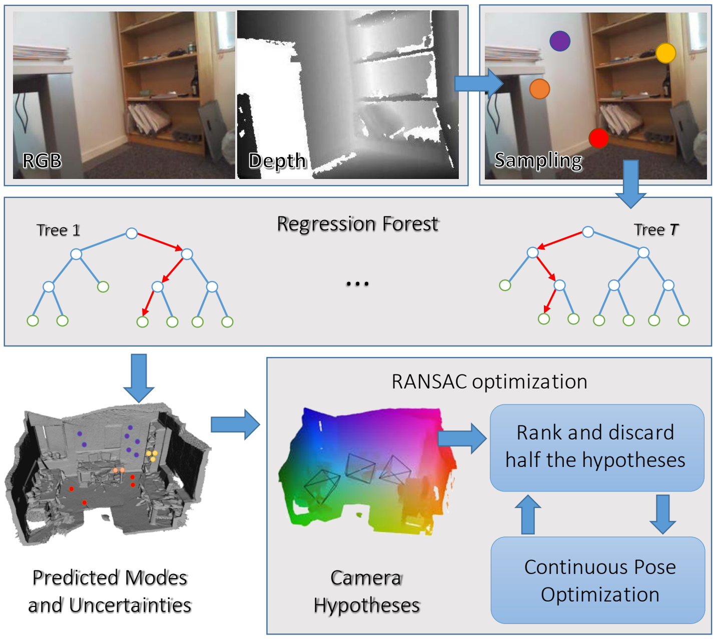
RGB
Depth
Sampling
Tree 1
Regression Forest
...
Tree T
RANSAC optimization
Rank and discard half the hypotheses
Continuous Pose Optimization
Predicted Modes and Uncertainties
Camera Hypotheses

Figure 1: **Camera Relocalization Pipeline.** As a new frame arrives (top left), pixels are sparsely sampled (top right) and passed down a regression forest (middle). The forest predicts a series of candidate locations in the scene as well as the uncertainty associated with each prediction (bottom left). Given these predictions, camera pose hypotheses are robustly sampled (bottom middle) and continuously optimized over iterations of RANSAC (bottom right).

In more detail:

- Given samples from the scene for which the relocalization task will be performed, a regression tree is trained using an objective function that minimizes the spatial variance of the scene coordinates. The distribution at the leaves is typically anisotropic and multimodal, and thus the leaves predict anisotropic Gaussian mixture models that have been fitted to those distributions.
- At test time, pixels are randomly sampled and passed down the regression forest. For each of these samples, the ensemble learner predicts a Gaussian mixture model that specifies a probability density function over that sample's location in the scene's 3D world coordinates. The predictions are then aggregated to generate robust camera hypotheses that are passed to a preemptive locally-optimized RANSAC. Each loop of the optimization ranks the camera hypotheses, discards the worst half, and optimizes the pose of all the remaining hypotheses by leveraging the predicted Gaussian mixtures. This process is repeated until only one hypothesis remains.

We present our main results and our comparison against the state of the art in Table 1. The proposed method yields major improvements over all baselines on all the test sequences. The decrease in relative error ranges from 18.7% on 'Stairs' to 90.7% on 'Heads', with a mean and median relative decrease of 50.3% and 46.0% respectively.

Scene	**Baselines**			**Our**
	Sparse RGB	[2]	[1]	**method**
Chess	70.7%	92.6%	96%	**99.4%**
Fire	49.9%	82.9%	90%	**94.6%**
Heads	67.6%	49.4%	56%	**95.9%**
Office	36.6%	74.9%	92%	**97.0%**
Pumpkin	21.3%	73.7%	80%	**85.1%**
RedKitchen	29.8%	71.8%	86%	**89.3%**
Stairs	9.2%	27.8%	55%	**63.4%**
Average	40.7%	67.6%	79.3%	**89.5%**

Table 1: **Main results on the 7-Scenes dataset.** Percentages denote the portion of frames that are below 5cm and 5° from the ground truth. Our method very substantially improves upon the baselines on all the scenes. Note that our results were generated using the same features across all sequences, while we compare against the best results from [1, 2] that use different features for each sequence ('DA-RGB' or 'DA-RGB + D').

In the paper, we illustrate that the empirical distribution of the samples is multi-modal and anisotropic at any level of the regression tree. This observation leads us to describing how to leverage that knowledge during the tree induction process and during the modeling of the predictions made in the leaves. Given a new test frame and the predictions made by the regression forest, we describe how we efficiently sample relevant camera candidates and detail how to rank those hypotheses using the predicted distribution of the pixels over the 3D scene coordinates. Finally, we show how to also leverage those distributions to efficiently optimize the pose of each hypothesis.

Our main conclusion is that modeling the distribution of each image pixel over the space of 3D scene coordinates and leveraging these uncertainties to optimize camera poses leads to camera relocalizations that are significantly more *robust* and *precise* than the state of the art.

[1] Abner Guzman-Rivera, Pushmeet Kohli, Ben Glocker, Jamie Shotton, Toby Sharp, Andrew Fitzgibbon, and Shahram Izadi. Multi-output learning for camera relocalization. In *Computer Vision and Pattern Recognition*. IEEE, 2014.

[2] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *Computer Vision and Pattern Recognition*. IEEE, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.