VS-Net: Voting with Segmentation for Visual Localization
- Supplementary Material -

Zhaoyang Huang1,2∗ Xiaowei Zhou1 Han Zhou1∗ Yijin Li1 Bangbang Yang1 Yan Xu2
Hujun Bao1 Guofeng Zhang1 Hongsheng Li2,3

1State Key Lab of CAD&CG, Zhejiang University
2CUHK-SenseTime Joint Laboratory, The Chinese University of Hong Kong
3School of CST, Xidian University

Abstract

In this document, we provide details of our implementation complementing the original paper. Furthermore, we inspect the performance of our scene-specific landmarks and the voting-by-segmentation algorithm with extended experiments. Lastly, a video showing practical localization performance of our framework is given.

1. Detail of Implementation

Landmark generation. SuperVoxel [6] is a 3D oversegmentation algorithm that uniformly sets vast seeds in 3D space and raises a patch for each pre-defined seeds. We set a seed resolution, which can be approximately regarded as the size of patches, in each scene and discards the seeds that do not cover surfaces. Therefore, we can generate more landmarks with a smaller patch size. We present the patch size and the corresponding number of landmarks we use in the visual localization experiment in Tab. 1. In contrast with SfM-based visual localization frameworks relying on the SfM map, our scene-specific landmarks are computed from a 3D surface. We compare the SfM feature map built by COLMAP [9] and the 3D surface we use reconstructed by

\begin{algorithm}
\caption{Landmarks from Votes}
\label{alg:landmarks}
\textbf{INPUT:}
\begin{itemize}
    \item \(p_i\): Pixel coordinates in current patch label
    \item \(d_i\): Pixel votes at \(p_i\)
    \item \(l_j^{(0)}\): Initial landmark of landmark \(j\) from RANSAC-based approach [7]
    \item \(\theta_{min}\): Threshold of minimum neighbor pixels
    \item \(\theta_{dist}\): Threshold of neighbor pixel distance
    \item \(\epsilon_{step}\): Epsilon of landmark coordinate change
\end{itemize}

\textbf{OUTPUT:}
\begin{itemize}
    \item \(l_{best}^j\): Best voting landmark
    \item \(b_{valid}\): Validity of the landmark
\end{itemize}

\begin{algorithmic}
    \STATE \(l_{best}^j \leftarrow l_j^{(0)}\)
    \STATE \(b_{valid} \leftarrow \text{true}\)
    \STATE \(t \leftarrow 1\)
    \WHILE{\(t < \text{Max Iteration}\)}
        \STATE \(S = \{||p_i - l_j^{(t-1)}|| < \theta_{dist}\}\)
        \IF{\(|S| < \theta_{min}\)} \text{false}
            \STATE \(b_{valid} \leftarrow \text{false}\)
            \STATE \text{break}
        \ENDIF
        \FOR{all \(p_i \in S\)}
            \STATE Compute normal of voting map \(n_i \leftarrow \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} d_i\)
        \ENDFOR
        \STATE \(l_j^{(t)} \leftarrow (\sum_S n_i n_i^\top)^{-1} (\sum_S n_i n_i^\top p_i)\)
        \IF{\(|l_j^{(t)} - l_j^{(t-1)}|| < \epsilon_{step}\)} \text{break}
        \ENDIF
        \STATE \(l_{best}^j \leftarrow l_j^{(t)}\)
        \STATE \(t \leftarrow t + 1\)
    \ENDWHILE
\end{algorithmic}
\end{algorithm}

1.1 Image Description

(a) SfM Map

(b) 3D Surface

Figure 1. Comparison of (a) SfM feature map and (b) TSDF surface.

\[ -85x-198 \]

\[ (a) \text{SfM Map} \quad (b) \text{3D Surface} \]

\[ -85x-198 \]
Table 1. The number of landmarks in the visual localization experiments.

<table>
<thead>
<tr>
<th>7Scenes</th>
<th>Cambridge Landmarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>College</td>
</tr>
<tr>
<td>Num.</td>
<td>4330</td>
</tr>
<tr>
<td></td>
<td>15cm</td>
</tr>
<tr>
<td></td>
<td>2.0m</td>
</tr>
<tr>
<td></td>
<td>5cm</td>
</tr>
<tr>
<td></td>
<td>13878</td>
</tr>
<tr>
<td></td>
<td>4081</td>
</tr>
</tbody>
</table>

Figure 2. Scene-specific landmarks vs. Reg vs. HSC-Net in individual scenes.

Figure 3. The reconstructed map (gray points) and camera trajectory (purple) of Street in the Cambridge Landmarks Dataset.

KinectFusion [5] in Fig. 1. The SfM feature map is typically messy and not accurate enough even in a static indoor scene, while the 3D surface is more tidy and accurate, which benefits the following localization.

VS-Net details. We use the DeepLabv3 [1] as the backbone for landmark segmentation and add a decoder for pixel-wise voting. Our models are trained by Adam optimizer with a learning rate of 1e-4, a weight decay of 5e-4, and a batch size of 2. The images are respectively scaled to $640 \times 480$ and $960 \times 540$ in the 7-Scenes dataset and the Cambridge Landmarks dataset. The length of prototype embedding used in the 7-Scenes dataset and the Cambridge Landmarks dataset are 24 and 64. $k$ is set as 2 in the $k$NN search of prototype-based triplet loss. We use affine transformation data augmentation in training with the same parameter settings in hsc-net [3] on both of these datasets. We elaborate the EM-like landmark location refinement algorithm in Alg. 1.

**Case study of Street.** There are six scenes in the Cambridge Landmarks Dataset. Following previous works [3], we only compare VS-Net with others in five scenes except Street because the provided poses in Street is not reasonable. We present a part of the 3D map and the camera trajectory computed from official ground-truth camera poses of the Street in Fig. 3. The cameras within 0.5m to the wall are rendered as green. As we can see, the camera trajectory drifts seriously.

2. Extended Experiments

Scene-specific landmarks vs. Reg vs. HSC-Net in individual scenes. We compare scene-specific landmarks and scene coordinates in individual scenes. Reg [3] is the base-
Figure 4. Feature matching with fundamental matrix based RANSAC.

In the view of feature-based visual localization, landmark labels can be regarded as the feature descriptors that associate two detected landmarks. To show that VS-Net generates more robust correspondences in a specified scene, we compare our landmark association with SIFT [4], D2-Net [2] feature matching in Figs. 4. All the matches have been filtered by epipolar line in scene coordinate regression and HSC-Net [3] is the state-of-the-art scene coordinate regression method. We compare VS-Net with them to show the superiority of our proposed scene-specific landmarks. In addition to the pose accuracy metrics, we measure the robustness and accuracy of correspondences by two other metrics: the outlier ratio and mean reprojection error of inliers. We compute the reprojection error of scene-specific landmarks and scene coordinates with ground-truth poses. The established 2D-to-3D correspondences whose reprojection errors are lower than 10 pixels are regarded as inliers, and the others are treated as outliers. In Fig. 2, the mean inlier reprojection error and outlier ratio of scene-specific landmarks are concentrated on lower error intervals, which means that the accuracy and robustness of scene-specific landmarks are better than scene coordinates in most query images. The number of poses computed from scene-specific landmarks consistently surpasses scene coordinates at the low threshold in both positional error and angular error, which indicates that scene-specific landmarks exhibit better pose estimation performance in most scenes.
geometry with RANSAC. R2D2 [8] is not able to figure out more than three correspondences, which can not be applied to the following fundamental matrix estimation, so we do not present it in the figures. The images are extracted from the Seq-3 in the 7-Scenes chess scene, which is a test sequence. The image pairs are constituted by the first frame and a frame sampled from the subsequent frames. Even after RANSAC, there are many erroneous matches in SIFT and D2-Net when the baseline is large because the textures are locally similar. We highlight some of them with red circles. By contrast, our landmark association keeps good robustness and accuracy under such a significant viewpoint changing.

Visual localization in practical environments We provide a supplementary video that qualitatively demonstrates the performance of VS-Net in severe practical environments. The Cambridge Landmarks dataset provides some challenging images and videos while they are not contained in the test set. We compute the camera poses of the images through the proposed visual localization framework and project the model into the views of the original sequences for performance evaluation. The successfully detected and associated landmarks are visualized with red crosses for better understanding. The results are presented in the supplementary video, which indicates the prominent robustness of our framework.

References


