**Time: from 05/04/2019 to 11/04/2019**

1. **Paper reading:**

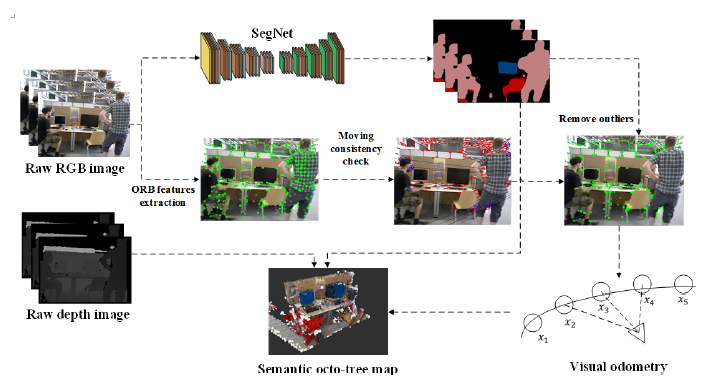
DS-SLAM: A Semantic Visual SLAM towards Dynamic Environments

Main Author：Chao Yu

Publication: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)

Main idea:

In this paper, a robust semantic visual SLAM towards dynamic environments named DS-SLAM is proposed. Semantic information is not only used to generate an octo-tree map based representation of the environment but also utilized to filter outliers in the process of tracking in a dynamic environment.



The raw RGB images captured by Kinect2 are processed in tracking thread and semantic segmentation thread simultaneously. The tracking thread first extracts ORB feature points, then check moving consistency of the feature points roughly and save the potential outliers. Then the tracking thread waits for the image that has pixel-wise semantic label predicted by semantic segmentation thread. After the segmentation result arrives, the ORB feature points outliers located in moving objects will be discarded based on the result and the potential outliers detected before. Then, the transformation matrix is calculated by matching the rest of stable feature points.

The idea of moving points detection in this paper is straightforward.

1. calculate optical flow pyramid to get the matched feature points in the current frame.
2. if the matched pair is too close to the edge of the image or the pixel difference of the 3×3 image block at the center of the matched pair is too large, the matched pair will be discarded.
3. find fundamental matrix by using RANSAC with the most inliers. Then calculate the epipolar line in the current frame by using the fundamental matrix.
4. determine whether the distance from a matched point to its corresponding epipolar line is less than a certain threshold. If the distance is greater than the threshold, then the matched point will be determined to be moving. The distance is determined as follows:

where, X Y Z represent line vector and F represents fundamental matrix

**Paper reading:**

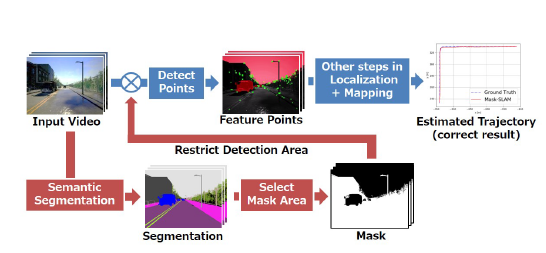
Mask-SLAM: Robust feature-based monocular SLAM by masking using semantic segmentation

Author: Masaya Kaneko

Publication: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops

Main idea:

In this paper, they propose a new framework to exclude feature points using a mask produced by semantic segmentation. Excluding feature points in masked areas enables vSLAM to stably estimate camera motion.



1. Randomly extract a sufficient number of samples fromthe data.

2. Estimate a set of parameters to fit these samples. This set of parameters is called hypothesis.

3. Apply the obtained hypotheses to all data other than the extracted samples and compute the distances between the estimated parameters for each data sample.

4. Consider samples with small distances as inliers and let the number of inliers represent the correctness of a hypothesis.

5. The above operation is performed a number of times and the hypothesis with the largest number of inliers is adopted, while outlier data is excluded.

**3. Code reading:**

http://webdii.unizar.es/~raulmur/orbslam

**Next week work plan**

Continue to read http://webdii.unizar.es/~raulmur/orbslam

Continue to do research on semantic segmentation SLAM

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Time: from 28/03/2019 to 04/04/2019**

**1.Paper reading:**

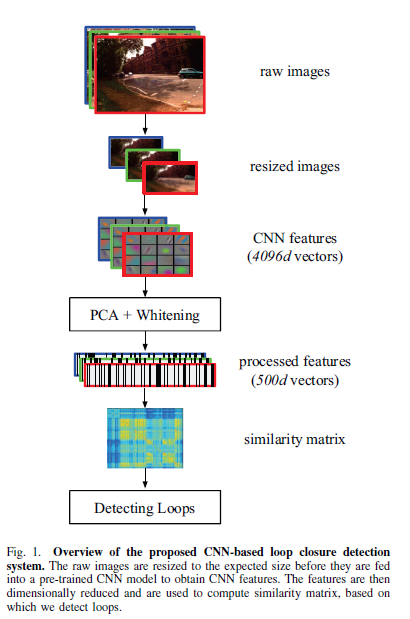
Loop Closure Detection for Visual SLAM Systems Using Convolutional Neural Network

Author: Xiwu Zhang

Publication: Proceedings of the 23rd International Conference on Automation & Computing, University of Huddersfield , Huddersfield, UK, 7-8 September 2017

Main idea:

We propose a loop closure detection method based on convolutional neural networks

****

1. Normalization: For each feature vector extracted from CNN model, we perform normalization step as follows:
2. PCA Dimensionality Reduction: Suppose we have obtained normalized feature vectors and the corresponding matrix 𝑋 consists of these vectors is:

a principal component analysis procedure is performed as the following algorithm:

Let

For i=1 to n do

Replace in X with

End for

1. Whitening: they whitened each feature vector according to the form:
2. Euclidean distance:
3. Similarity score :

If the similarity score is larger than a specific threshold, we regard it as a loop

**2. Do research on semantic segmentation SLAM**

**3. Code reading:**

Github.com/gaoxiang12/slambook

**Next week work plan**

Continue to read Github.com/gaoxiang12/slambook

Continue to do research on semantic segmentation SLAM

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Time: from 21/03/2019 to 27/03/2019**

**Paper reading：**

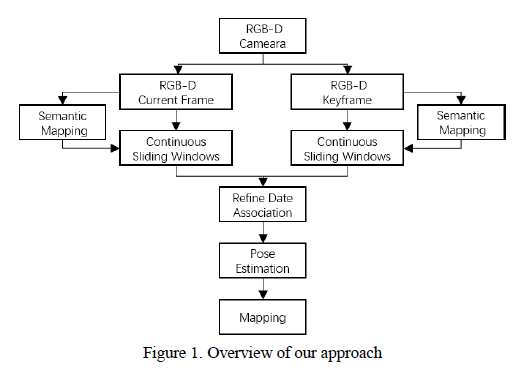
Robust RGB-D SLAM in Dynamic Environment Using Faster R-CNN

Author: Sifan Yang

Publication：2017 3rd IEEE International Conference on Computer and Communication

**Main idea:**

This paper proposed a method which is used in dynamic environment. We first check whether there are objects moving by the threshold which represents consistency of matching and identify every potential candidate of the dynamic object. If the dynamic existence is confirmed, we compute the dynamic region and figure out the dynamic object efficiently. Then we will cull the wrong data association in dynamic region and add more new data association we don’t have in static region..



1. Use Faster-RCNN to detect and identify the potential candidate of the dynamic object whose category will be labeled.

2. Distinguish the stationary from the dynamic environment and refine the data association by removing the mismatching related to the dynamics.

They compute the consistency of the point J in corresponding object regions.

It means that we project the feature points in the stationary region of the current frame k into the corresponding region of the keyframe h with the estimated camera pose . Then they compute the similarity between these. If the J is over the threshold, they label the region of object as dynamic status from stationary status and update the data association by filtering out the data associations in the moving region. Once the status is dynamic, there’s no chance for status to turn back. If the J is within the threshold, they label the region as stationary state and reserve previous data association.

3. Estimate the camera pose with better data association and the optimization of the graph.

4. With accurate pose estimation, they reconstruct the dynamic environment successfully.

关于语义分割的SLAM的调研感悟

SLAM的另一个大方向就是和深度学习技术结合。到目前为止，SLAM的方案都处于特征点或者像素的层级。关于这些特征点或像素到底来自于什么东西，我们一无所知。这使得计算机视觉中的SLAM与我们人类的做法不怎么相似，至少我们自己从来看不到特征点，也不会去根据特征点判断自身的运动方向。 我们看到的是一个个物体，通过左右眼判断它们的远近，然后基于它们在图像当中的运动推测相机的移动。之前，研究者就试图将物体信息结合到SLAM中，曾把物体识别与视觉SLAM结合起来，构建带物体标签的地图。另外，把标签信息引入到BA或优化端的目标函数和约束中，我们可以结合特征点的位置与标签信息进行优化。语义信息可以帮助SLAM提高建图和定位的精度，特别是对于复杂的动态场景。传统SLAM的建图和定位多是基于像素级别的几何匹配。借助语义信息，我们可以将数据关联从传统的像素级别升级到物体级别，提升复杂场景下的精度。借助SLAM技术计算出物体之间的位置约束，可以对同一物体在不同角度。不同时刻的识别结果进行一致性约束，从而提高语义理解的精度。 综合来说，SLAM和语义的结合点主要是以下方面：

传统的物体识别、分割算法往往只考虑一幅图，而在SLAM中我们拥有一台移动的相机。如果我们把运动过程中的图片都带上物体标签，就能得到一个带有标签的地图。另外，物体信息亦可为回环检测、BA优化带来更多的条件。

**Code reading:**

Github.com/gaoxiang12/slambook

**Next week work plan**

Continue to read Github.com/gaoxiang12/slambook

Read paper: CNN&SLAM

**Time: from 13/03/2019 to 20/03/2019**

**Paper reading：**

CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction

Author：Keisuke Tatento

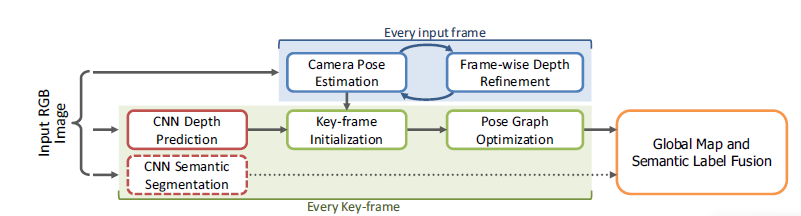
Publication: CVPR2017

**Main idea:**

This paper investigates how predicted depth maps from a deep neural network can be deployed for accurate and dense monocular reconstruction.

**Framework**

The flow diagram in Fig. 1 sketches the pipeline of the framework. Authors employ a key-frame based SLAM paradigm. Within such approach, a subset of visually distinct frames is collected as key-frames, whose pose is subject to global refinement based on pose graph optimization. At the same time, camera pose estimation is carried out at each input frame, by estimating the transformation between the frame and its nearest key-frame. To maintain a high frame-rate, they propose to predict a depth map via CNN only on key-frames. In particular, if the currently estimated pose is far from that of existing keyframes, a new key-frame is created out of the current frame and its depth estimated via CNN.



**CNN model for SLAM**

The depth prediction architecture is based on ResNet50 and initialized with pre-trained weights on ImageNet. Pooling and FC are replaced by a sequence of residual up-sampling blocks composed of a combination of unpooling and convolutional layers. After up-pooling, drop-out is applied. The loss function is based on the **reverse Huber function.**

They also retrained this network for predicting pixel-wise semantic labels for RGB images. In this way, they modified the network so that it has as many output channels as the number of categories and employed a soft-max layer and a cross-entropy loss function to be minimized via back-propagation and SGD.

**Key-frame Creation & Pose Graph Optimization**

There is a problem that sensors for SLAM have different intrinsic parameters from those used to capture the training set, the results will be inaccurate. In this way, they propose to adjust the depth regressed via CNN with the ratio between the focal length of current camera, and that of the sensor used for training, as

Where is the depth map directly regressed by CNN

This transformation is estimated by minimizing the photometric residual between the intensity image of the current frame and the intensity image of the nearest key-frame via weighted Gauss-Newton optimization based on the objective function

Where is Huber norm and is a function measuring the residual uncertainty. And r is the photometric residual defined as

while represents a 3D element of the vertex map computed from the key-frame’s depth map

Once is obtained, the current camera pose in the world coordinate system is computed as

作者首先筛选出关键帧，在关键帧上用训练好的CNN网络来预测单帧图深度值得到深度图，并以此深度图作为SLAM架构先验深度。同时在关键帧上用训练好的另一个CNN网络来做语义分割。

随后像直接法SLAM的一样做BA，用高斯牛顿法，基于pose graph方法优化得到pose，和普通的半稠密 SLAM过程基本一样。

将深度图和语义分割图融合进全局已有的场景深度图（实际上是三维地图点集合了）和三维语义分割图中

**Code reading:**

Github.com/gaoxiang12/slambook

**Next week work plan**

Continue to read Github.com/gaoxiang12/slambook

Read paper: Fully Convolutional Networks for Semantic Segmentation