Project Update 1 - Motion Segmentation

Dieker, Jonas and Hiebl, Christian

Informatics - Technische Universität München

Introduction

The literature review section is just for completeness sake and our own records. Feel free to skip that section. The main idea is briefly introduced in the last section. We also prepared some further questions to be discussed in our meeting.

Literature Review

Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation [1]

Competitive collaboration is an unsupervised three player setup consisting of two neural nets (static scene reconstructor, moving region reconstructor) competing for training data that is regulated by a third neural net, moderator (motion segmentation network).

Architecture: Two networks R (depth, camera motion $\rightarrow$ reasons about static regions) and F (optical flow $\rightarrow$ reasons about moving regions) compete for a resource (data) and the third network M (moderator) distributes data to R and F. In the second step R and F collaborate to update M.

Input: Sequence of RGB images

Output: Monocular Depth & Camera Motion $\rightarrow$ Optical flow of static regions, Optical Flow Estimation, Motion Segmentation; Final: Composite Optical Flow

MonoRec: Semi-Supervised Dense Reconstruction in Dynamic Environments From a Single Moving Camera [2]

Semi-supervised monocular dense reconstruction for depth maps from a single camera. Dense 3D reconstruction can be split into two main directions: Multi-view stereo (MVS) method and monocular depth prediction methods. The former have been improved by CNNs in the recent years, assume a static environment and generalize better to monocular depth prediction methods. These rely fully on deep learning and are able to reconstruct moving objects.

Architecture: Siamese U-Net - Multiple RGB images with their respective camera poses to output motion segmentation (MaskModule) and depth map (DepthModule). In the MaskModule, semantic information from a pretrained ResNet18 as well as geometric priors from the cost volumes between frame I_t and $I_{t'}$. The DepthModule outputs a dense inverse depth map in which the predicted mask is used to deliminate wrong depth prediction for moving objects. Supervised sparse depth loss obtained by visual odometry; self supervised photometric loss; no manual labelling or LiDAR depth for DepthModule.

Input: Assembled cost volume - a tensor that is the aggregated photometric consistency

Output: Dense inverse depth map

Learning To Segment Rigid Motions From Two Frames [3]

Independent object motions can be recovered from an egomotion field. Its takes two consecutive frames as input and predicts segmentation masks for the background and multiple rigidly moving objects, which are then parameterised by 3D rigid transformations.

Idea that detectors should rely less heavily on appearance-based cues but rather on geometric constraints. There are fundamental difficulties that plague geometric motion segmentation → construct motion cost maps to address motion degeneracies.

Architecture U-Net for motion segmentation (predicts binary pixel-wise labels), CenterNet for instance segmentation mask using the head proposed in PolarMask; supervised;

Input: Consecutive RGB images

Output: Rigid Body Scene Flow (Egomotion, rotation translation of each rigid body)

Motion-based Object Segmentation based on Dense RGB-D Scene Flow [4]

Proposing learned methods from raw data over multiple scales instead of traditional methods which rely on brightness constancy, smooth motion within a small region and thus cannot deal with non Lambertian surfaces, occlusions or large displacements.

Architecture: Encoder-Decoder style network. Two Siamese networks take pairs of consecutive RGB and depth images. The paper uses the correlation layer of FlowNetC for RGB encoding and to associate encoded point cloud features.

Input: Consecutive RGB-D images (split as RGB and XYZ - Siamese neural network)

Output: Object scene flow and motion based rigid object segmentation (from decoder which predicts object position, rotation, translation)

Mask R-CNN [5]

Framework for object instance segmentation = detection + semantic segmentation. It extends Faster R-CNN by adding another branch in the head for predicting semantic segmentation masks for each RoI proposed by the RPN (Region Proposal Network). To fix misalignment due to the RoIPool, which rounds image sizes after convolutions, RoIAlign is introduced.

Architecture: very similar to Faster R-CNN. Same backbone. In the head an extra branch for the mask prediction is introduced. Note: head of the network is applied to each generated RoI.

Input: One RGB image

Output: BBox, mask

Neural Network Architecture

The network takes two successive RGB frames as its input. The two images are then fed through two branches in parallel. The first of which is an off-the-shelf network M with which we obtain the scene flow and the second being a Mask R-CNN. We then use the instance mask to obtain the scene flow for each instance detected and subsequently fit and decompose essential matrices from the flow maps to find R and T for each instance. Finally, we compare the transformation of each instance with the ego-motion to classify it as being a static or dynamic object, from this the motion segmentation mask is produced.

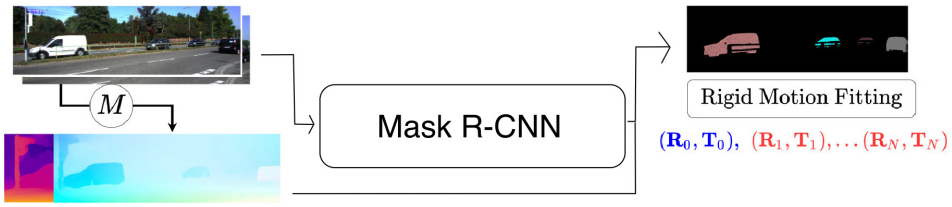

Figure 1: Proposed Initial Architecture (adapted from [4])

Self-supervised: Uses its own predictions for the ground truth; more viable in our situation. How to adapt network to be self-supervised?

References

[1] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J. Black. "Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 12240–12249.

[2] Felix Wimbauer, Nan Yang, Lukas von Stumberg, Niclas Zeller, and Daniel Cremers. "MonoRec: Semi-Supervised Dense Reconstruction in Dynamic Environments From a Single Moving Camera". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 6112–6122.

[3] Gengshan Yang and Deva Ramanan. "Learning To Segment Rigid Motions From Two Frames". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 1266–1275.

[4] Lin Shao, Parth Shah, Vikranth Dwaracherla, and Jeannette Bohg. "Motion-Based Object Segmentation Based on Dense RGB-D Scene Flow". In: *IEEE Robotics Autom. Lett.* 3.4 (2018), pp. 3797–3804.

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. "Mask R-CNN". In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2980–2988.