Scientific report

on project implementation status between July-December 2018

Proiect PN-III-P4-ID-PCCF-2016-0180

Financing: UEFISCDI, Nr. 9/2018

Project homepage: http://www.imar.ro/clvp/projects.php?ID_p=15

Integrated Semantic Visual Perception and Control for Autonomous Systems

"Simion Stoilow" Institute of Mathematics of the Romanian Academy (IMAR)

Principal Investigator (PI): Prof-univ. dr. Cristian Sminchisescu

Co-PI: Prof-univ. Dr. Sergiu Nedevschi

During the first six months of the project we have been involved in starting out the activities for project PN-III-P4-ID-PCCF-2016-0180 (acronym SEPCA). The project goes well along the guidelines set out in the plan of action, in partnership with the research group of Prof. Sergiu Nedevschi from the Technical University of Cluj-Napoca. We have assembled a research team composed out of experienced researchers (V. Olaru, PhD, University of Karlsruhe 2004) and young, but experienced researchers (A. Zanfir, E. Marinoiu, M. Zanfir, A. Costea, Robert Varga, Andra Petrovai, Vlad Miclea), who got their master degrees under the guidance of the experienced researchers in the two groups at IMAR and UTCN (Olaru, Oniga, Nedevschi, Sminchisescu). The individual activities has started right away according to the work plan (a more detailed description of them, including results, follows below). Also, a workshop has been organized at IMAR to initiate the collaboration in the project between 30-31.10.2018.

A1. Deep 3D Reconstruction

Task 1.1: Deep Learning of Graph Matching under Global Constraints

The problem of graph matching under node and pairwise constraints is fundamental in areas as diverse as combinatorial optimization, machine learning or computer vision, where representing both the relations between nodes and their neighborhood structure is essential. In an article published at CVPR 2018, we present an end-to-end model that makes it possible to learn all parameters of the graph matching process, including the unary and

pairwise node neighborhoods, represented as deep feature extraction hierarchies. The challenge is in the formulation of the different matrix computation layers of the model in a way that enables the consistent, efficient propagation of gradients in the complete pipeline from the loss function, through the combinatorial optimization layer solving the matching problem, and the feature extraction hierarchy. Our computer vision experiments and ablation studies on challenging datasets like PASCAL VOC keypoints, Sintel and CUB show that matching models refined end-to-end are superior to counterparts based on feature hierarchies trained for other problems. For qualitative results, fig. T1.1.1 below shows matching results obtained on the PASCAL VOC dataset. Qualitative results on the CUB dataset, including ground truth results are shown in fig. T1.1.2.

Fig. T1.1.1 Twelve qualitative examples of our best performing network on the PASCAL VOC test-set. For every pair of examples, the left shows the source image and the right the target. Colors identify the computed assignments between points. The method finds matches even under extreme appearance and pose changes.

Fig. T1.1.2 Four qualitative examples of our best performing network on the CUB-200-2011 test-set. Images with a black contour represent the source, whereas images with a red contour represent targets. Color-coded correspondences are found by our method. The green framed images show ground-truth correspondences. The colors of the drawn circular markers uniquely identify 15 semantic keypoints.

Reference

A. Zanfir and C. Sminchisescu. “Deep Learning of Graph Matching”, Proceedings - 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018

Task 1.2: Deep Structured Geometric Models with Semantics

Towards the goals we developed MubyNet – a feed-forward, multitask, bottom up system for the integrated localization, as well as 3d pose and shape estimation, of multiple people in monocular images. The challenge is the formal modeling of the problem that intrinsically requires discrete and continuous computation, e.g. grouping people vs. predicting 3d pose. The model identifies human body structures (joints and limbs) in images, groups them based on 2d and 3d information fused using learned scoring functions, and optimally aggregates such responses into partial or complete 3d human skeleton hypotheses under kinematic tree constraints, but without knowing in advance the number of people in the scene and their visibility relations. We design a multi-task deep neural network with differentiable stages where the person grouping problem is formulated as an integer program based on learned body part scores parameterized by both 2d and 3d information. This avoids suboptimality resulting from separate 2d and 3d reasoning, with grouping performed based on the combined representation. The final stage of 3d pose and shape prediction is based on a learned attention process where information from different human body parts is optimally integrated. State-of-the-art results are obtained in large scale datasets like Human3.6M and Panoptic, and qualitatively by reconstructing the 3d shape and pose of multiple people, under occlusion, in difficult monocular images. Qualitative results are shown below in fig. T1.2.1.

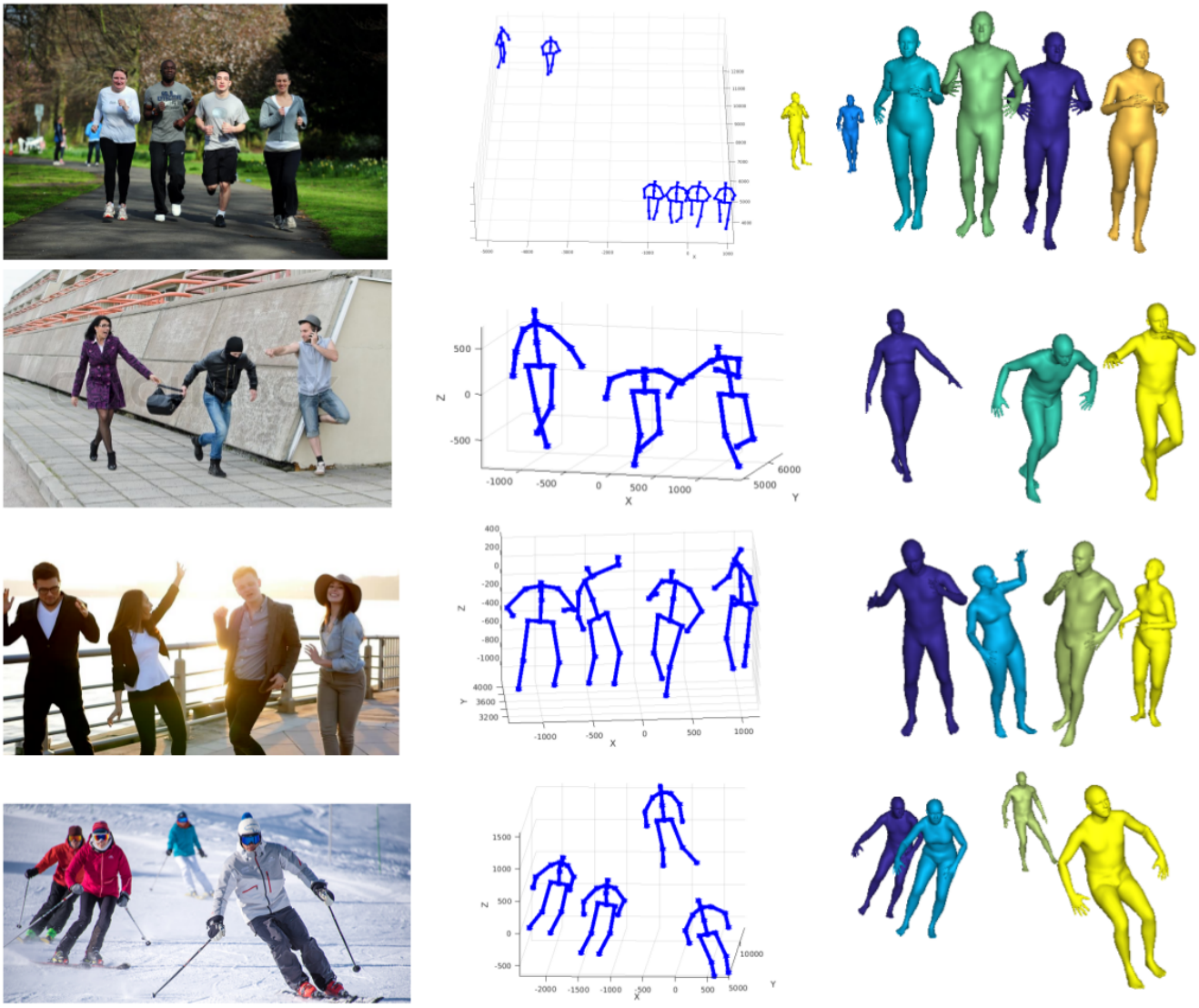

Fig T1.2.1 Human pose and shape reconstruction of multiple people produced by MubyNet illustrate good 3d estimates for distant people, complex poses or occlusion. For global translations, we optimize the Euclidean loss between the 2d joint detections and the projections predicted by our 3d models.

Reference

A. Zanfir, E. Marinoiu (Oneata), M. Zanfir, A. Popa, C. Sminschisescu. “Deep Network for the Integrated 3D Sensing of Multiple People in Natural Images”, Proceedings of the Thirty-second Conference on Neural Information Processing Systems, NIPS 2018

Task 1.2 deals with producing a 3D representation of the environment, in which each 3D point is also associated with semantic class. A contribution with respect to this goal “Real-Time Semantic Segmentation-Based Stereo Reconstruction” was done by generating the depth map (first step required for the 3D representation) by enhancing the stereo reconstruction process with semantic information. To this end, initially a semantic map of the scene is generated by using a convolutional neural network. Then,

each sub-task of the stereo reconstruction algorithm is tailored to incorporate scene information obtained from the semantic map and thus to enhance the results. New learning algorithms (based on genetic algorithms and convolutional neural networks) are introduced for these steps. Results show that the new method produced the best real-time stereo reconstruction results on the Kitti stereo benchmark. Although using stereo reconstruction on aerial images is both cumbersome and susceptible to errors (the system can easily decalibrate), this approach is really important because it demonstrates the benefits of using high-level scene information (provided through the semantic map) for low-level vision tasks required for depth perception.

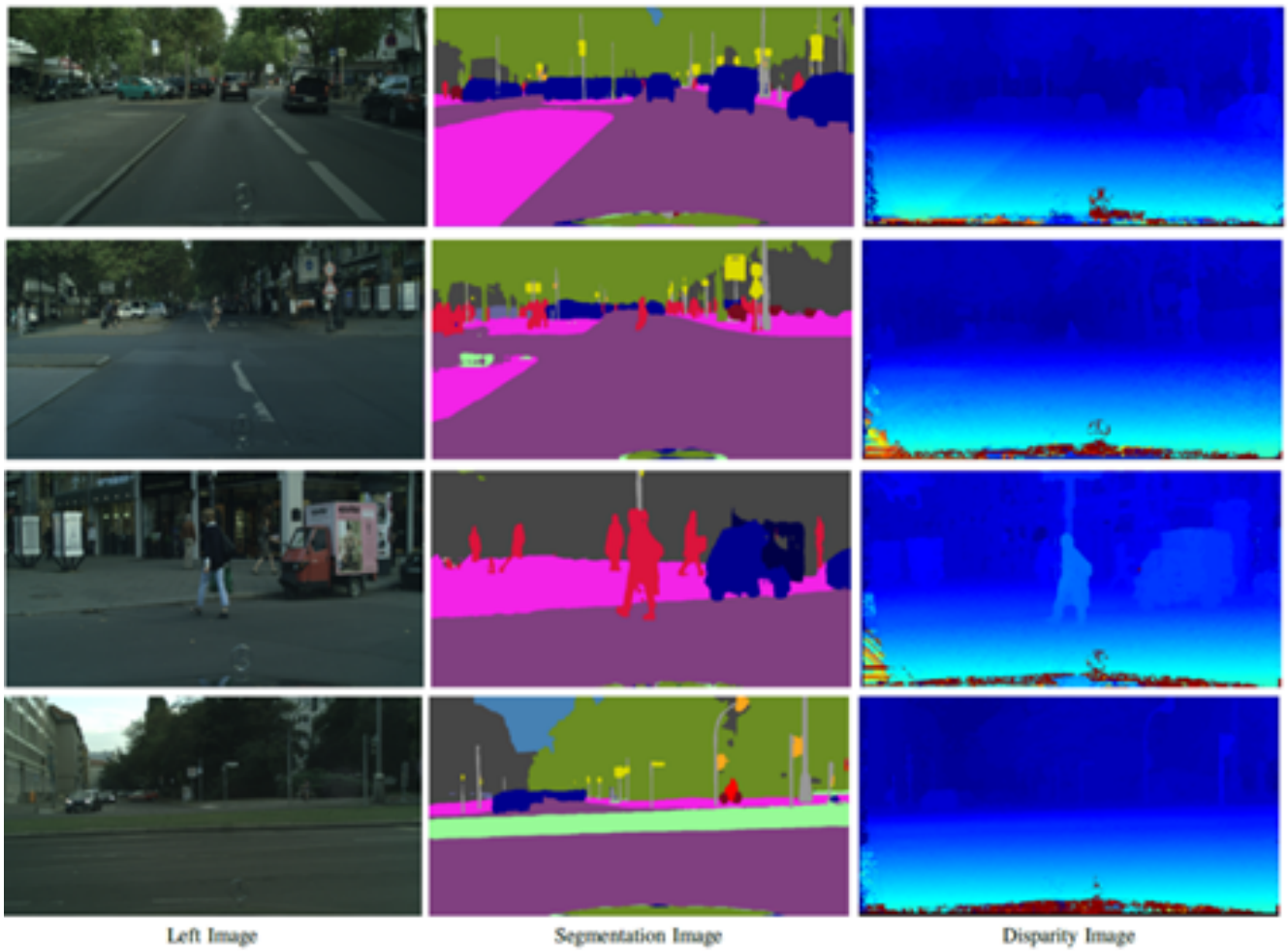

Fig. T1.2.2 Results from the paper “Real-Time Semantic Segmentation-Based Stereo Reconstruction”. The disparity image (from Cityscapes dataset) is obtained by enhancing the stereo pipeline with information from the semantic segmentation. The proper shape of the objects is clearly captured in all scenarios

Reference

V. Miclea, S. Nedevschi, “Real-Time Semantic Segmentation-Based Stereo Reconstruction”, IEEE Transactions on Intelligent Transportation Systems, accepted for publication

A2. Visual Recognition and Localization

Task 2.2. Active and adversarial learning structures and methods for dynamic data.

This task focuses on the design and of computational procedures that are amenable to the large-scale training of dynamic data. During the course of the project, we studied, designed and implemented novel convolutional architectures for panoptic image segmentation. Panoptic segmentation provides pixel-level classification and instance identifiers for dynamic objects in the scene.

In our first solution ”Fusion Scheme for Semantic and Instance-Level Segmentation”, we propose a fusion scheme for instance and semantic segmentation based on heuristics to solve conflicts, which could be applied as a fast post-processing step on top of any semantic and instance network. We base our fusion module on the observation that semantic segmentation performs well in background segmentation, but struggles with foreground classes, where pixels of objects having different class but same category (for example truck, bus) are often misclassified (Fig. 1, column 3). In the case of instance segmentation, pixels of objects are correctly classified but the mask is more coarse than in the case of segmentation. We propose a post-processing step that provides a correction mechanism by propagating instance class and label on the semantic path at category level. We observe significant accuracy increases especially in the case of large objects.

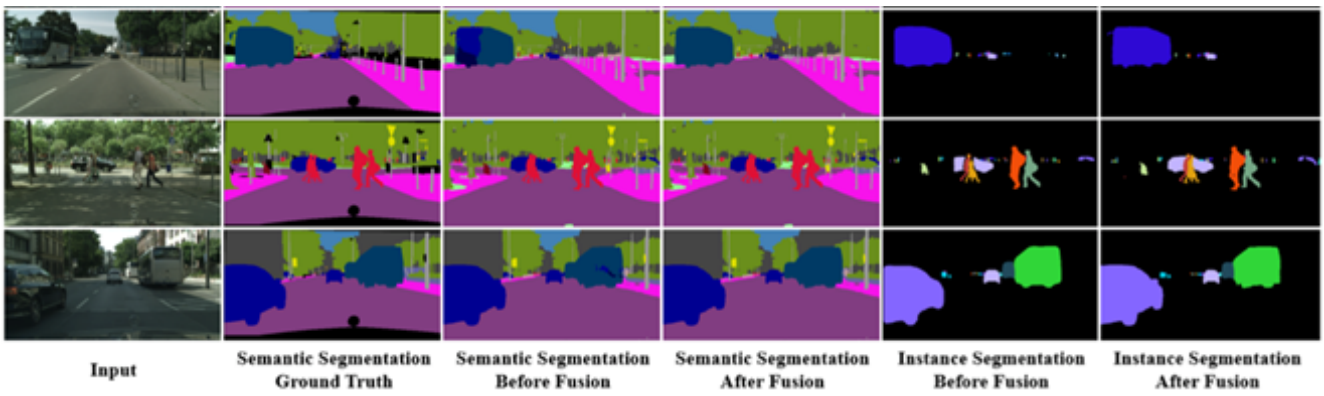

Fig. T2.2.1 Results from the paper ”Fusion Scheme for Semantic and Instance-Level Segmentation”. The proposed post-processing fusion scheme yields a more accurate semantic segmentation, especially in the case of large objects.

Reference

A. D. Costea, A. Petrovai, S. Nedevschi, ”Fusion Scheme for Semantic and Instance-Level Segmentation”, Proceedings of 2018 IEEE Intelligent Transportation Systems Conference (ITSC), Maui, Hawaii, USA, November 4-7, 2018.