Machine Perception 2021 - Project report

Optical Flow Estimation

Maxime Raafat
raafatm@student.ethz.ch

Nicolas Muntwlyer
municola@student.ethz.ch

ABSTRACT

Human optical flow plays an important role in the analysis of human action. We present a deep learning based approach to predict the optical flow in multi-human scenes. Our method is based on PWC-Net, a general optical flow estimation network. We use the Multi-human optical flow dataset to fine-tune the before-mentioned model, and introducing an iterative refinement procedure and a cyclic loss to achieve a significant improvement in the flow computation. We furthermore introduce a novel architecture which runs in harmony with PWC-Net and makes additional use of pre-computed segmentation masks.

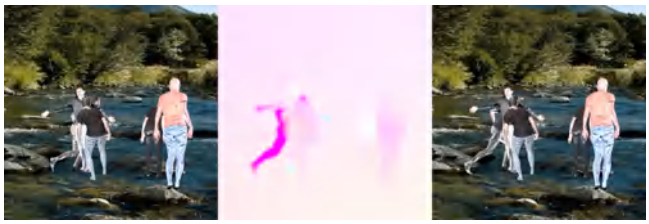

Figure 1: Example of an optical flow prediction (middle) between Image 1 (left) and and Image 2 (right)

1 INTRODUCTION

Predicting an accurate optical flow plays a core part in many computer vision applications such as autonomous driving, video compression and video editing. For the past years classical methods have strongly dominated the field, while most methods implement an energy minimization pioneered by Horn and Schunck [3]. However recent advances in machine learning and deep learning have led to innovative learning-based approaches[8, 10] that not only outperform classical methods, but also make flow computation significantly faster and enable real-time applications.

Predicting an accurate optical flow is a challenging task to solve : since flow estimation requires per-pixel localization, a deep learning model not only needs to learn relevant features from a scene frame, but also requires to match those features between two input images. A further challenge is the increasing demand for real-time performance for applications on mobile device.

While PWC-Net [10] achieves impressive generalizability to predict optical flow, many applications care specifically for human flow estimation. Ranjan et al. [9] therefore generated a *Multi-human optical flow* (MHOF) dataset, which we use to fine-tune PWC-Net on human scenes.

In this work we fine-tune a PWC-Net-based architecture on the MHOF dataset and present an iterative refinement procedure (similar to IRR-NET [5]) together with a newly introduced cycle-loss to further increase the flow prediction accuracy. Lastly we explore an approach combining the intermediate flow-predictions of PWC-Net with a novel network, Seg-Net, inspired from LiteFlowNet [4] and which makes additional use of the segmentation masks of the two input images.

2 RELATED WORK

In many areas of computer vision, powerful deep learning models are progressively replacing classical methods. For optical flow estimation it started with the pioneering work of Dosovitskiy et al. [1]. Although they could not achieve state-of-the-art results, they showed that CNN's have high potential for flow prediction tasks.

By stacking multiple FlowNet modules on top of each other, Ilg et al. [6] developed FlowNet2, competing on par with state-of-the-art methods. Although Flownet2 achieves astonishing results, its size is large and does not yet outperform classical approaches. Hui et al. [4] came up with a network based on FlowNet2 while being 30 times smaller in model size. By introducing a cascaded flow inference relying on sub-pixel refinement (which will be the foundation for our novel architecture Seg-Net), LiteFlowNet not only outperforms FlowNet2 in size, but is also 1.36 times faster in running speed.

With PWC-Net, Sun et al. [10] presented a new simplistic model with far less parameters than Flownet2 [6] and significantly outperforming it. PWC-Net is based on the well-established principles of pyramidal processing, warping and the use of a cost volume. By reducing PWC-Net to only one pyramid layer, which is repeatedly called to iteratively update the final flow through the addition of the individual residual flows, IRR-Net [5] further reduces the number of parameters while not losing in prediction accuracy.

Creating datasets for supervised optical flow learning is a challenging task which is simplified if data is synthesized such that ground truth is known. While large scale datasets like FlyingChairs focus on general optical flow prediction, MHOF [9] is first to introduce a dataset tailored for human motion prediction. Other approaches for self-supervised learning [7, 11] have also been developed in order to remove the need for ground truth optical flow generation procedures, however our work builds on PWC-Net, which requires ground truth samples.

3 METHOD

3.1 Problem Statement

We solve the following task: given two input images I_1 and I_2, predict the optical flow $\mathbf{w}$ between the two images. As additional input we get a body segmentation mask for both images and we are provided with the PWC-Net model pre-trained on the FlyingChairs dataset. The images contain multi-human scenes with arbitrary backgrounds. We measure the model accuracy with the average End Point Error (EPE) over all image pixels.