Evolution of optical flow estimation with deep networks

October 28, 2020

1 Introduction

FlowNet managed to demonstrate that optical flow estimation can be considered as a learning problem. The aim of paper [2] is to propose another network able to improve performances of FlowNet especially on small displacements and real word data. In FPAR project is useful because paper [1] exploits optical flow by warping it. On top of that, optical flow lonely can be considered as a possible variation.

2 Main concepts

The new proposed network (called FlowNet 2.0) has important modifications in view to be a feasible model also in real-world applications such as action recognition and motion segmentation. FlowNet 2.0 is built with 3 ideas.

- **Influence of dataset schedules** : Usage of multiple datasets improves significantly results. So, the focus is on the importance of training data.
- **Warping operation**: Development of a stacked architecture that includes the warping of the second image with intermediate optical flow.
- **Subpixel motion**: Application on small displacements and real-world data; a special training dataset and a specialized on small motions network are created.

To have a better generalization, so in view to reach optimal performances on arbitrary displacements, a further network is added. Its goal is to learn how to fuse the former stacked network with the small displacement network. The final network performs definitevely better compared to the initial FlowNet. Moreover, FlowNet 2.0 follows the strategy called **curriculum learning**, which consists of training machine learning models on a series of gradually increasing tasks. Of course in this case it is applied on computer vision tasks.

3 Dataset schedules

As said before, high quality training data is crucial for the success of a classification task. Actually, it was discovered that not only the kind of data is important, but also the order in which the datasets are inserted during training phase, especially if the sources have different properties. In this case study, two datasets were analyzed.

- **Flying chairs dataset**: 22k image pairs of chairs, imposed on random background images from Flickr. In it, some random affine transformations are applied on the chairs and on the background in view to obtain the second image and ground truth flow fields. So, the dataset contains only planar motions.
- **Flying things 3D**: A 3D version of the previous dataset. So, in addition to the Flying chairs dataset, there are motion 3D, lighting effects and a more variety among the object models.

The authors discovered that training the network immediately on the 3D dataset (of course more realistic) leads to poor results, while the network, before trained on the Chairs dataset and then on Things, is more efficient. The cause might be that the Chairs dataset helps the network to learn the general concept of color matching, avoiding possible confusing priors for 3D motion and realistic light too early. Finally, the fine tuning is performed on 3D Things dataset in view to catch more difficult details.

4 Stacking two networks

The aim was to test if also deep networks could benefit from an iterative approach, because all the optical flow estimation strategies were based on iterative methods. The experiment consists of stacking two networks.

- The first network takes as inputs images I_1 and I_2.
- Subsequent networks get I_1, I_2 and the previous flow estimate w_i=$(u_i, v_i)^T$, where i represents the index of the network in the stack.
- To have a trace of the previous errors and to compute an incremental update easier for the network, they also **optionally** warp the second image $I_2(x, y)$ throught the flow w_i and bilinear interpolation to $I_{2,i}(x, y) = I_2(x + u_i, y + v_i)$. In this way the next network in the stack can focus on the remaining increment between I_1 and $I_{2,i}$. When use warp, also the error e_i is provided to the network, where $e_i = ||I_{2,i} - I_1$. Due to the fact that bilinear interpolation is performed, the derivatives of the warping operation can be computed; as a consequence the training of stacked networks is enabled.

As a conclusion of these trials, it was discovered that a simple stacking network without warping leads to overfitting. On the other side, through the inclusion of warping operation, stacking always improves results and also the presence of an intermediate loss after the first network is very convenient to the network. For all these reasons, best results were achieved when keeping the first network fixed and only training the second network after the warping operation.

5 Small displacements

The last issue is to make the network also able to recognize small displacements. To achieve that, the whole network stack was trained on a mixture of datasets in view to have more generalization as possible. In addition, also a non-linearity is applied to the error in view to not losing performances in large displacements. It was sufficient for small displacements, but not yet for subpixel motion, because the noise remains a problem. To try overcoming this issues these operations were tried:

- Stride 2 of the first layer removed;

- Kernels 7X7 and 5X5 changed in multiple 3X3 kernels
- Addition of convolutions between the upconvolutions to obtain smoother estimates, because noise is especially a problem with small displacements.

So, keeping in mind all these points, a small final network was built. This network has as inputs the flows, the flow magnitudes and the errors in brightness after warping. On top of that, it contacts the resolution two times by a factor 2 and expands again.

References

[1] *ego RNN's reference paper at.* URL: https://arxiv.org/pdf/1807.11794.pdf.

[2] *Optical flow evolution reference paper at.* URL: https://arxiv.org/pdf/1612.01925.pdf.