2018 17th International Symposium on Distributed Computing and Applications for Business Engineering and Science

MaxFlow: a Convolutional Neural Network Based Optical Flow Algorithm for Large Displacement Estimation

Yufen Sun, Gang Liu
School of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
Email: yufen@whut.edu.cn, liu_gang@whut.edu.cn

Lei Xie
Intelligent Transport Systems Research Center
Wuhan University of Technology
Wuhan, China
Email: xielei@whut.edu.cn

***Abstract*—Optical flow estimation is a basic problem in computer vision. FlowNet is the first convolutional neural network based optical algorithm that estimates optical flow by learning the relationship between image pair and the corresponding optical flow. In this paper, MaxFlow is proposed to improve the accuracy of FlowNet. The architecture of MaxFlow is similar to that of FlowNetSimple. MaxFlow uses two kinds of new layers, which are designed specially for estimating large displacements of small scale objects. The new down sampling layer makes the network to predict the maximum displacement in a region. Thus the large movements will not be missed. The new up sampling layer up samples the estimated optical flow fields without using any parameter. It simplifies the network without decreasing the accuracy of the network. Experiments on synthetic datasets and real datasets illustrate that the two new layers are effective and the accuracy of MaxFlow is higher than that of FlowNet.**

***Keywords*-optical flow; convolutional neural networks; variational methods**

I. Introduction

Convolutional neural networks (CNN) have been successfully applied in many computer vision tasks. However, there are relatively few CNN-based optical flow algorithms. FlowNet [1] is the first algorithm that trains a CNN to estimate optical flow in an end-to-end manner. FlowNet 2.0 [2] improves FlowNet by adopting a more complex learning schedule and integrating multiple networks. In this paper, we improve FlowNet by giving it different supervision.

Estimating large displacements and small displacements can be thought as two different subproblems [3], and large displacement estimation has long be thought as the hard one [4], especially the large displacements of small scale structures [5]. This paper puts focus on estimating the large displacements of small scale structure.

FlowNet contains a contracting part that aggregates motion information and an expanding part that refines the coarse optical flow prediction. Neurons in FlowNet estimate optical flow by matching features. If the distance between two features is larger than the size of the receptive field of a neuron, the neuron cannot match them because it only sees features located in its receptive field. As a result, large displacements can only be outputted by neurons in the upper layers that have larger receptive fields. However, in FlowNet, the prediction target of each neuron in the upper layers is the weighted average of ground truth optical flow values in a large region. Thus, the fast motion of small scale structures will be lost if its scale is much smaller than the region. In lower layers, the region used for computing the prediction target of each neuron is smaller, but the receptive field is also smaller. FlowNet concatenates the coarse feature maps with the fine feature maps to enlarge the receptive fields of neurons in lower layers. These coarse feature maps can provide information for estimating large displacements, but it is hard for upper layers to learn features for predicting optical flow average and large displacements of small scale structures at the same time. In this paper, we simplify the task of upper layers and only require them to predict large displacements.

This paper proposes MaxFlow, a optical flow estimation CNN whose network structure is similar to that of FlowNet. In MaxFlow, the prediction target of each neuron in upper layers is the maximal displacement of the pixels in a region, not the average displacement. When the region reduces to one pixel, the network outputs the estimated optical flow field. MaxFlow can predict large displacements of small scale objects, and the optical flow field estimated by MaxFlow has more clear edge than that of FlowNet.

In the remainder of this paper, Section II introduces the related work. Section III introduces the new layers in MaxFlow. Section IV reports the experimental analysis and last, Section V gives the conclusion.

II. Related Work

Optical flow algorithms estimate the displacement of each pixel in two consecutive video frames. Since Horn and Schunck published their seminal paper in 1981 [6], variational methods have become the dominating methods for optical flow estimation. These methods are effective at estimating small displacements. Using coarse-to-fine warping schemes, algorithms can estimate large displacements of large scale structures by computing optical flow at coarser resolution levels [4], but they may fail to estimate the large displacements of small scale structures. Brox and Malik introduced descriptor matching to variational methods to resolve the difficulties in large displacement estimation [5]. The sparse matches computed

978-1-5386-7445-1/18/$31.00 ©2018 IEEE
DOI 10.1109/DCABES.2018.00040

by descriptor matching provide the information of large displacements for variational methods.

EpicFlow [7] makes a further step to utilize the information provided by descriptor matching. On the sparse matches computed by DeepMatching [8], EpicFlow performs edge-preserving interpolation to get a dense flow field, which is a good initialization for the optimization of variational models. In this way, the large displacement information provided by sparse matching and the sub-pixel accuracy provided by variational methods are combined effectively. After the publication of EpicFlow, matching, interpolation and variational refinement become three standard steps in a modern optical flow algorithm [9]. Many algorithms put the focus on inventing effective descriptor matching techniques suited for optical flow estimation [10]–[13]. Some algorithms use CNNs to generate image patch descriptors for matching [3], [11], [12], [14]–[16]. Besides sparse matching techniques, discrete optimization can also be used to estimate large displacements. DiscreteFlow [17] and FullFlow [18] perform discrete inference to estimate large-displacement integral optical flow. The obtained integral displacements are also refined by EpicFlow.

However, even for the state-of-the-art optical flow algorithms that use descriptor matching to estimate large displacements, there is still much space to improve the accuracy for large displacement estimation. Yang and Soatto design a method [19] that runs Flow Fields algorithm [10] multiple iterations. At each iteration, only the estimations with high confidence are determined. They find that large displacements of small scale structures are determined lasted.

FlowNet [1] is the first algorithm that uses CNN to estimate dense optical flow directly. It adopts a fully convolutional structure [20], so the input images could be of any size. The contracting part of FlowNet uses convolutional layers to extract features, and the expanding part uses upconvolutional layers to upsample the coarse feature maps and coarse optical flow estimations. The upsampled feature maps and estimations are concatenated with fine feature maps to estimate fine optical flow fields. The network training loss is the weighted sum of the endpoint errors (EPE) of all optical flow estimations. Compared with EpicFlow, FlowNet preserves more fast motion details, but it makes some errors in small background movements. Variational methods can be used to improve the estimations of FlowNet further. FlowNet 2.0 [2] approximates the variational refinement by stacking multiple FlowNets, and uses a small network specially trained for small optical flow to improve the estimation accuracy. We find that though FlowNet preserves more motion details, some large displacements of small scale structures are missing in the optical flow fields. We try to solve this problem by requiring the network to estimate the maximum displacement of pixels in each region.

III. The Network

The architecture of MaxFlow is similar to that of FlowNetSimple [1]. The contracting part of MaxFlow contains ten convolutional layers, each followed by a ReLU nonlinearity layer. The expanding part of MaxFlow contains six groups of layers for optical flow refinement. In each refinement step, the up sampled coarse flow estimation, the up sampled coarse feature maps, and the corresponding fine features are concatenated. Then the fine flow estimation is estimated by performing convolution on this concatenation. The expanding part predicts seven optical flow fields with different resolutions. Each prediction is compared with a down sampled ground truth to get a loss.

The main differences of FlowNet and MaxFlow lie in the expanding part. In MaxFlow, the down sampling layers used to down sample the optical flow ground truths and the up sampling layers used to up sample the optical flow predictions are different from those in FlowNet. These two kinds of new layers put emphasis on large displacement discovery.

A. The Down Sampling Layer

As FlowNet, MaxFlow computes optical flow estimations of different resolutions. Each estimation is inputted into a loss layer to compare with the ground truth. The resolutions of these estimations are smaller than the ground truth. FlowNet down samples the ground truth by computing the weighted sum of the flow vectors in a region. For each estimation, the size of the region equals $(2n+1)\times(2n+1)$ if the resolution of the flow estimation is n times smaller than that of the ground truth. This down sampling is reasonable for regions in which the flows are smooth. However, when there are different movements in the region, the physical meaning of the weighted sum is unclear. More seriously, the large displacements of small or thin objects will be smoothed out.

The down sampling layer in MaxFlow does not compute the weighted sum of the flow vectors, but chooses the maximum absolute value for each dimension of the flow as the ground truth for a region. This ground truth makes the network to predict the maximum displacement in a region. Thus, the large displacements will be predicted even for small scale structures. When the region reduces to one pixel, the network predicts the optical flow for each pixel. We adopt $n \times n$ as the size of the region for down sampling, so the region reduces to one pixel when the resolution of the flow estimation equals that of the ground truth.

In order to obtain optical flow estimation whose resolution equals the resolution of the input, MaxFlow performs two more steps of estimation refinement than FlowNet. FlowNet does not refine the optical flow estimation to the input resolution because compared to bilinear up sampling, further refinement does not significantly improve the results [1]. Thus, the improvements of MaxFlow do not come from the added layers for refinement.

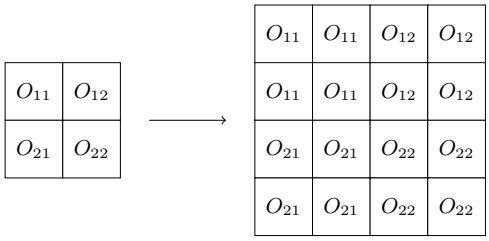


Figure 1. The up sampling operation.

B. The Up Sampling Layer

Both FlowNet and MaxFlow refine a coarse flow estimation by performing convolution on the concatenation of up sampled coarse feature maps, fine features, and the up sampled coarse flow estimation. FlowNet up samples both the coarse feature maps and the coarse flow estimation using an 'upconvolutional' layer ('Deconvolution' layer in Caffe [21]). This layer first enlarges the feature maps by padding 0, then applies convolution to the enlarged feature maps.

When convolution is applied to a coarse flow estimation, the flow estimation is changed. To transfer the information of large displacements to later layers, we prefer to keep the coarse flow estimation unchanged. Thus, we realize a new up sampling layer that just copies the flow vectors from the coarse estimation to the enlarged feature maps. This up sampling operation on one feature map is illustrated in Fig. 1. As in the down sampling layer, one pixel in the coarse estimation corresponds to a region in the fine feature maps. The size of the region is $n \times n$ if the resolution of the coarse flow estimation is n times smaller than the enlarged feature maps. For every pixel in a region in the enlarged feature maps, its flow value equals the flow value of the pixel in the coarse flow estimation that corresponds to this region. This new up sampling layer has no parameter, which reduces the flexibility of the network. However, the preserved information of large motion is beneficial to the estimation for large displacements.

IV. Experiments

We use the Flying Chairs dataset [1] to train our MaxFlow network. The experimental analysis is performed on Flying Chairs, KITTI, and Sintel datasets. We compare the MaxFlow with FlowNet [1]. All of our experiments are performed on an Intel Xeon E5 at 2.4GHz with a Nvidia GTX 1080.

A. Training Details

The network is trained by the synthetic Flying Chairs dataset. This dataset contains 22,872 image pairs, among which 22,232 image pairs are used to train the network, and the remain 640 image pairs are used to test the network.

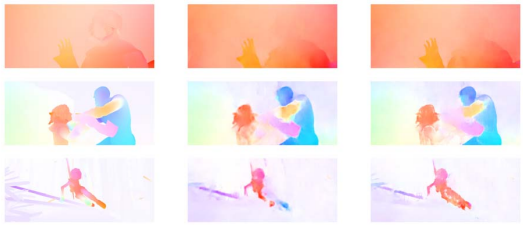
We use the same training process as that used by FlowNet [1]. The training loss of the network is the average endpoint error, which is the average Euclidean distance between the predicted optical flow vectors and the corresponding ground truths. The Adam optimization algorithm is used, with a mini-batch size of 8. The learning rate starts from 0.0001 and is divided by 2 every 100k iterations after the first 300k. The network is trained for 600k iterations. On our computer, the whole training process takes about three days.

Table I
THE ACCURACY COMPARISON.

Models	Flying Chairs Test	Sintel Clean (train)	KITTI 2012
FlowNetS	2.71	4.50	8.26
MaxFlow-	2.49	4.50	8.11
MaxFlow	2.51	4.43	8.10

B. Accuracy Comparison

The accuracy of the networks are measured by the average endpoint errors of the predicted flow values. Table I gives the average endpoint errors of FlowNet and MaxFlow on three different datasets. The errors of MaxFlow are lower than that of FlowNetS. Figure 2 shows the optical flow fields estimated by these two networks on the Sintel Clean (train) dataset. Figure 3 shows the estimated optical flow fields on the Flying Chairs dataset. We can see that compared with FlowNetS, our MaxFlow preserves more motion details and have more clear edges.



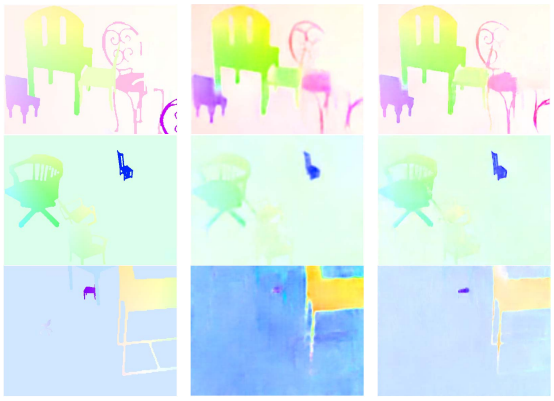
(a) (b) (c)

Figure 2. The accuracy comparison on the Sintel Clean (train) dataset. (a) The ground truth. (b) The optical flow predicted by FlowNetS. (c) The optical flow predicted by MaxFlow.

There are two kinds of new layers in MaxFlow. In order to investigate the effects of the new layers, we also give the average endpoint errors of MaxFlow-, in which only the new down sampling is used, in Table I. From the data in Table I, we can see that the down sampling layers improve the accuracy of the network, and the up sampling layers simplify the network without decreasing the accuracy of the network.

V. Conclusion

This paper proposes MaxFlow, a new convolutional network for optical flow estimation. The key contribution is the two new layers designed specially for estimating large displacements of small scale structures. Experiments illustrate that MaxFlow can preserve more motion details than FlowNet, and has lower average endpoint error on multiple datasets.



(a) (b) (c)

Figure 3. The accuracy comparison on the Flying Chairs dataset. (a) The ground truth. (b) The optical flow predicted by FlowNetS. (c) The optical flow predicted by MaxFlow.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under grant 51479158.

REFERENCES

[1] A. Dosovitskiy, P. Fischery, E. Ilg, P. Husser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 2758–2766, .

[2] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1647–1655, .

[3] T. Schuster, L. Wolf, and D. Gadot, "Optical flow requires multiple strategies (but only one network)," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, .

[4] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Computer Vision - ECCV 2004*, T. Pajdla and J. Matas, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 25–36.

[5] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 500–513, March 2011, .

[6] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, no. 1-3, pp. 185–203, 1981, . [Online]. Available: http://dx.doi.org/10.1016/0004-3702(81)90024-2

[7] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Epicflow: Edge-preserving interpolation of correspondences for optical flow," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1164–1172, .

[8] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Deepflow: Large displacement optical flow with deep matching," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 1385–1392, .

[9] S. Zweig and L. Wolf, "Interponet, a brain inspired neural network for optical flow dense interpolation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, , ł.

[10] C. Bailer, B. Taetz, and D. Stricker, "Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 4015–4023, .

[11] J. Xu, R. Ranftl, and V. Koltun, "Accurate optical flow via direct cost volume processing," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, .

[12] C. Bailer, K. Varanasi, and D. Stricker, "Cnn-based patch matching for optical flow with thresholded hinge embedding loss," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, .

[13] Y. Hu, R. Song, and Y. Li, "Efficient coarse-to-fine patchmatch for large displacement optical flow," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, .

[14] F. Gney and A. Geiger, "Deep discrete flow," in *Asian Conference on Computer Vision (ACCV)*, 2016, .

[15] J. Thewlis, S. Zheng, P. H. S. Torr, and A. Vedaldi, "Fully-trainable deep matching," *BMVC*, vol. abs/1609.03532, 2016, . [Online]. Available: http://arxiv.org/abs/1609.03532

[16] D. Gadot and L. Wolf, "Patchbatch: A batch augmented loss for optical flow," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 4236–4245, .

[17] M. Menze, C. Heipke, and A. Geiger, "Discrete optimization for optical flow," in *Pattern Recognition*, J. Gall, P. Gehler, and B. Leibe, Eds. Cham: Springer International Publishing, 2015, pp. 16–28.

[18] Q. Chen and V. Koltun, "Full flow: Optical flow estimation by global optimization over regular grids," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, .

[19] Y. Yang and S. Soatto, "S2f: Slow-to-fast interpolator flow," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, , .

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3431–3440, , .

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 675–678, . [Online]. Available: http://doi.acm.org/10.1145/2647868.2654889