Integrated Variational And Nearest Neighbor field (IVANN) for Optical Flow

Zhuoyuan Chen
Baidu Research
1195 Bordeaux Dr, Sunnyvale, CA
chenzhuoyuan@baidu.com

Ying Wu
Northwestern University
2145 Sheridan Road, Evanston, IL
yingwu@eecs.northwestern.edu

Hailin Jin, Zhe Lin and Scott Cohen
Adobe Systems Inc.
345 Park Ave, San Jose
{hljin,zlin,scohen}@adobe.com

Abstract

It is a fundamental problem to construct accurate dense correspondences between two images. Despite the efforts and promising methods handling relatively small motion, one remaining challenge is induced by large and complex non-rigid motion. Aiming at this challenge, the new method proposed exploits the mutual boosting between the variational flow and the nearest-neighbor field (NNF). The proposed method "IVANN" gives a very effective solution under rather complex motion, and currently achieved state-of-the-art performance on both the Middlebury[3] and MPI-Sintel benchmarks[7].

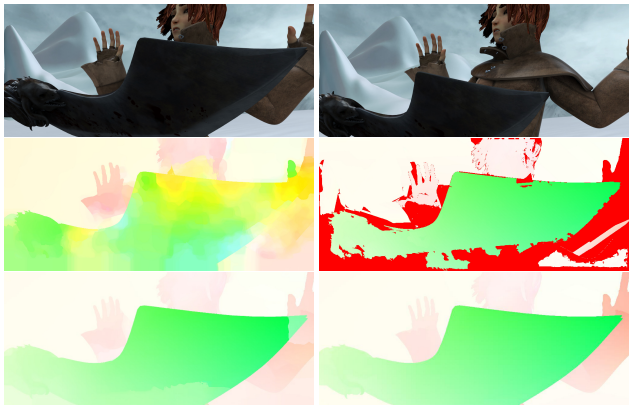

Figure 1: **Top left, right**: Frame 14 and 15 from the Sequence "Ambush 2" [7]; **Middle left**: color-coded motion from the variational flow [17]; **Middle Right**: color-coded motion from the NNF Matching [5]. We color incorrect motion with red. **Bottom left**: our final result. **Bottom right**: ground truth.

1 Introduction

Inferring a dense motion field between two images is one of the most fundamental problems. It can be dated back to the early 80s with the original seminal work [12]. There have since been many great advances [19, 8], as indicated by the Middlebury benchmark [3]. However, a good solution still remains elusive in challenging situations such as complex large displacement motions. This paper addresses particularly the issue in optical flow.

Most existing methods are based on linearizing the optical flow constraint. However, their performance highly depends on the quality of initial motion field, while large complex scenarios makes the numerical optimization prone to low quality local optima. In the absence of any prior knowledge, **zero** are used as the initialization, which is then refined by a gradient-based optimization technique. These methods can only recover small deviations around the initial value. To handle large deviations, most methods adopt a multi-scale framework, with the insight that motion at lower resolution is generally smaller. Sub-sampling reduces the size and motion within, but the reduction in image size leads to a loss of motion details unrecoverable. Therefore, these methods perform poorly on image structures with motions larger than their size, which is an intrinsic limitation of the traditional framework.

In this paper, we propose to incorporate a different type of correspondence information between two images, namely *nearest neighbor fields* or *NNF* [5]. An NNF is defined as, for each patch in one image, the most similar patch in the other image. While computing exact NNF can be computationally expensive, a very recent attempt [5, 13] provides very efficient approximate solutions based on its strong spatial dependency.

1.1 Related Work

There is a huge body of literature on optical flow following the original work of Horn and Schunck [12]. We only discuss the papers that address the large displacement motion problem, since it is the main focus of this work.

The coarse-to-fine framework was first proposed in [2]. It has since been adopted by most optical flow algorithms to handle large motions. [1] was probably the first to note that the standard coarse-to-fine framework may not be sufficient. Brox and Malik [6] proposed to add robust keypoint matching with SIFT [14] into the optical flow framework to handle arbitrarily large motion with no performance sacrifice. Xu et al. [19] also proposed to use similar techniques but with fusion. Both [6] and [19] are based on keypoint detection and matching algorithms and may suffer in regions with weak texture due to lack of reliable keypoints.

Our work is closely related to [8, 11, 18, 6, 19, 4] combining feature matching with dense registration, especially to [8, 11, 18, 4] where NNF is also applied. The method in [8] assumes that the dominant motion patterns are well-behaved statistically globally, while [15, 11, 18, 4] denoise motion locally by either edge-aware filtering [15, 4], clustering [11] or max-pooling [18]. It is worth mention that Deep Matching[18] relies on dense HOG descriptors (Histogram of Gradient Orientations), which is embedded in a top-down convolution-pooling framework.

1.2 Contributions

The main contribution of our work is a high-accuracy optical flow framework that can handle large displacement motions. In particular we improve upon existing methods in the following ways:

- We use approximate NNF algorithms to initialize the dense correspondence field. It contains a high percentage of approximately accurate motions to recover the dominant motion patterns.
- We propose an efficient approach to handle large displacement. Observing that NNF and traditional flow have complementary power, we formulate the flow estimation as a motion segmentation problem by combining NNF and traditional flow results.
- Experimentally, our algorithm achieved a top ranking on the Middlebury [3] and MPI-Sintel [7] benchmarks.

Our IVANN Algorithm
1. Construct image pyramids, set scale $l = 0$,
initialize $\mathbf{u}_l = 0$
2. Propagate $\mathbf{u}_l$ to level $1 + 1$
3. Flow Refinement:
3.1 Continuous variational refinement
3.2 Denoised NNF
3.3 Adaptively Fuse results from 3.1 and 3.2 to obtain $\mathbf{u}_l$
4. If l is not the finest scale, go to step 2.

Table 1: The proposed IVANN algorithm.

2 Our Solution– the IVANN Method

We formulate our general motion estimation problems as an optimization problem with the objective function:

$$E(\mathbf{u}) = E_D(\mathbf{u}) + E_S(\mathbf{u}) \tag{1}$$

where $E_D(\mathbf{u})$ measures the matching error or data penalty, and $E_S(\mathbf{u})$ regularizes the flow field. We base our penalty function on the l_1-norm to reject outliers and preserve motion boundaries [19].

To optimize 1 is not trivial. We propose to use a denoised NNF with [11] as well as a traditional optical flow for numerical optimization. Then, we leverage an adaptive fusion method to obtain our final solution. The whole framework is shown in Figure 2.

2.1 Denoised-NNF

Given a pair of input images, we first compute an approximate NNF between them using PatchMatch [5]. As empirically studied in [8], the NNF is approximately consistent with the ground truth flow field. Accordingly, we apply the Non-Rigid Dense Correspondence (NRDC) [11] to aggregate the NNF since it is more robust to complex motion.

2.2 Traditional Variational Flow

We also apply the off-the-shelf classical optical flow methods to obtain an alternative solution. Specifically, we use the codes [17], which contain many modern techniques to achieve good performance, such as a coarse-to-fine framework, warping, robust cost function and so forth.

2.3 NNF Meets Variational Flow

While the denoised NNF enjoys more flexibility and robustness than the traditional variational flow, it also brings a lot of noises, especially vulnerable in case of large smooth regions, repetitive patterns and occlusions. However, traditional optical flow methods generally work well. Accord-

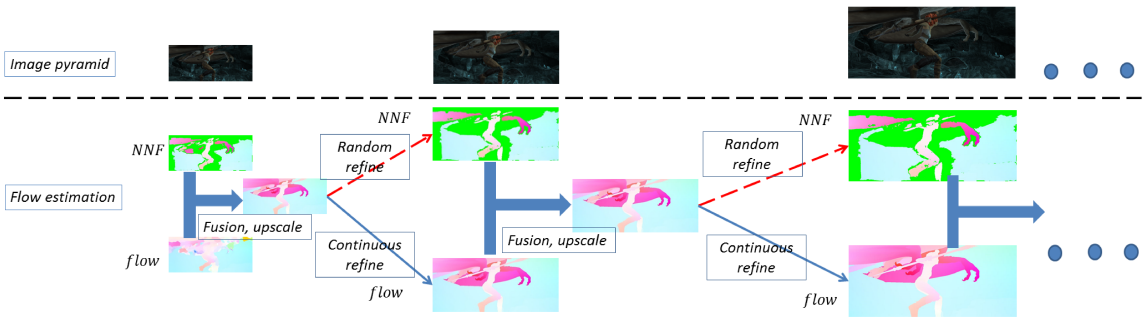

Figure 2: An illustration of our approach: variational flow and NNF are integrated in a multi-scale framework.

Figure 3: Complex Motion on real sequences[13]. The first and second columns are the input frames. The third column shows that NNF-Local captures the motion of the fast moving objects in these examples.

ingly, we also obtain a result from variational flow [17] and adaptively fuse with denoised-NNF.

Given the two flow fields: $\mathbf{u}_1$ from denoised NNF and $\mathbf{u}_2$ from variational flow, our final step is an adaptive fusion. We formulate the integration of two alternatives as a binary labeling problem:

$$\begin{aligned} \mathbf{u}^* &= \arg\min_{\mathbf{u}} E(\mathbf{u}) = E_D(\mathbf{u}) + E_S(\mathbf{u}) \\ s.t. &\quad \mathbf{u}(\mathbf{x}) \in \{\mathbf{u}_1(\mathbf{x}), \mathbf{u}_2(\mathbf{x})\} \end{aligned} \tag{2}$$

and then apply a QPBO fusion [16].

3 Experiments

Qualitative Evaluation

To further evaluate our approach, we apply NNF-Local on some real sequences with complex motions from VidPair dataset [13], namely the Sequence "Jackass", "Prince of Persia" and "Resident Evil Afterlife". Some results are shown in Figure 3. As we can see, NNF-Local can capture the fast moving objects quite accurately and the results are visually pleasant.

Quantitative Evaluation

Finally, we quantitatively evaluate our algorithm on the MPI-Sintel [7] and Middlebury [3] benchmarks.

For the Middlebury benchmark, we listed the Average End-point Error and Average Angle Error (AAE) of our algorithm in Table 3. At the time of publishing, our method achieves state-of-the-art quantitative results on the benchmark. In Figure 4, we copied the ranking from the evaluation websites [7]. Our method is ranked top at submission.

On the large-scale MPI-Sintel benchmark [7], we obtained an EPE of 7.249 and 5.386 on the final and clean video frames respectively. Our methods ranks 20-th among the 70 submissions. Our results are comparable to the state-of-art result of 5.459 and 3.102. It is noticeable that our algorithm achieves fairly small "EPE matched" on Sintel, indicating its ability to perform very well on pixels with correspondences.

4 Conclusion

In this work, we make an attempt to tackle the dense correspondence problem in complex motion scenarios. We propose the "IVANN" approach, where the denoised NNF are interleaved with the variational flow in a top-down framework. Our experiments on MPI-Sintel and Middlebury benchmarks clearly show that our approach can achieve satisfactory performance. Furthermore, we notice a recent work of Deep-Flow [18, 9] uses a deep-learning framework to obtain high-quality large motion. Our next step will focus on leveraging deep learning for further improvement.

	Army	Mequon	Schefflera	Wooden	Grove	Urban	osemite	Teddy
AEPE	0.07	0.15	0.18	0.10	0.41	0.23	0.10	0.34
AAE	2.89	2.10	2.27	1.58	2.35	1.89	2.43	1.01

Table 2: Experimental results of our algorithm on Middlebury test set [3].

Average endpoint error	avg. rank	Army (Hidden texture) GT im0 im1 all	disc	untext	Mequon (Hidden texture) GT im0 im1 all	disc	untext	Schefflera (Hidden texture) GT im0 im1 all	disc	untext	Wooden (Hidden texture) GT im0 im1 all	disc	untext
IVANN [92]	2.6	0.07 1	0.20 2	0.05 1	0.15 1	0.51 3	0.12 5	0.18 1	0.37 1	0.14 1	0.10 2	0.49 3	0.06 2
OFLAF [80]	6.6	0.08 7	0.21 3	0.06 5	0.16 5	0.53 4	0.12 5	0.19 2	0.37 1	0.14 1	0.14 7	0.77 21	0.07 4
MDP-Flow2 [69]	7.5	0.08 7	0.21 3	0.07 14	0.15 1	0.48 1	0.11 1	0.20 4	0.40 4	0.14 1	0.15 17	0.80 26	0.08 11
NN-field [72]	8.3	0.08 7	0.22 13	0.05 1	0.17 7	0.55 6	0.13 10	0.19 2	0.39 3	0.15 6	0.09 1	0.48 2	0.05 1
Epistemic [82]	9.6	0.07 1	0.21 3	0.05 1	0.16 5	0.55 6	0.12 5	0.20 4	0.44 7	0.15 6	0.11 3	0.65 6	0.06 2
TC/T-Flow [79]	14.3	0.07 1	0.21 3	0.05 1	0.19 13	0.68 24	0.12 5	0.28 17	0.66 22	0.14 1	0.14 7	0.86 31	0.07 4
Layers++ [37]	15.6	0.08 7	0.21 3	0.07 14	0.19 13	0.56 8	0.17 24	0.20 4	0.40 4	0.18 17	0.13 6	0.58 4	0.07 4
ADF [66]	15.7	0.08 7	0.22 13	0.06 5	0.18 9	0.62 15	0.14 14	0.29 22	0.71 27	0.17 12	0.16 28	0.91 39	0.07 4
LME [71]	15.9	0.08 7	0.22 13	0.06 5	0.15 1	0.49 2	0.11 1	0.30 26	0.64 17	0.31 63	0.15 17	0.78 23	0.09 23
IROF++ [58]	16.5	0.08 7	0.23 18	0.07 14	0.21 24	0.68 24	0.17 24	0.28 17	0.63 16	0.19 27	0.15 17	0.73 17	0.09 23

Average endpoint error	Grove (Synthetic) GT im0 im1 all	disc	untext	Urban (Synthetic) GT im0 im1 all	disc	untext	Yosemite (Synthetic) GT im0 im1 all	disc	untext	Teddy (Stereo) GT im0 im1 all	disc	untext
IVANN [92]	0.41 1	0.61 1	0.21 2	0.23 2	0.66 2	0.19 1	0.10 4	0.12 8	0.17 11	0.34 1	0.80 4	0.23 2
OFLAF [80]	0.51 4	0.78 5	0.25 3	0.31 4	0.76 3	0.25 6	0.11 10	0.12 8	0.21 29	0.42 7	0.78 2	0.63 12
MDP-Flow2 [69]	0.63 13	0.93 14	0.43 14	0.26 3	0.76 3	0.23 5	0.11 10	0.12 8	0.17 11	0.38 3	0.79 3	0.44 4
NN-field [72]	0.41 1	0.61 1	0.20 1	0.52 42	0.64 1	0.26 9	0.13 30	0.13 26	0.20 22	0.35 2	0.83 5	0.21 1
Epistemic [82]	0.71 26	1.07 30	0.53 27	0.32 6	1.06 18	0.28 11	0.11 10	0.13 26	0.15 7	0.41 6	0.88 9	0.54 5
TC/T-Flow [79]	0.67 22	0.98 21	0.49 23	0.22 1	0.82 5	0.19 1	0.11 10	0.11 1	0.30 66	0.50 24	1.02 26	0.64 15
Layers++ [37]	0.48 3	0.70 3	0.33 6	0.47 29	1.01 13	0.33 30	0.15 49	0.14 44	0.24 41	0.46 13	0.88 9	0.72 26
ADF [66]	0.69 24	1.03 24	0.47 18	0.43 18	0.91 8	0.28 11	0.12 20	0.12 8	0.20 22	0.43 8	0.88 9	0.63 12
LME [71]	0.66 18	0.96 18	0.53 27	0.33 7	1.18 28	0.28 11	0.12 20	0.12 8	0.18 15	0.44 9	0.91 12	0.61 10
IROF++ [58]	0.60 10	0.89 10	0.42 13	0.43 18	1.08 21	0.31 21	0.10 4	0.12 8	0.12 4	0.47 15	0.98 20	0.68 23

Figure 4: AEPE on the the Middlebury [3] test set.

References

[1] Luis Alvarez, Joachim Weickert, and Javier Sanchez. Reliable estimation of dense optical flow fields with large displacements. *IJCV*, 39(1):41–56, 2000.

[2] Anandan. A computational framework and an algorithm for the measurement of visual motion. *IJCV*, 2(3):283–310, 1989.

[3] Simon Baker, Daniel Scharstein, J. P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 92(1):1–31, 2011.

[4] Linchao Bao, Qingxiong Yang, and Hailin Jin. Fast edge-preserving patchmatch for large displacement optical flow. *CVPR*, 2014.

[5] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM SIGGRAPH*, 24, 2009.

[6] Thomas Brox and Jitendra Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *PAMI*, 33(3):500–513, 2011.

[7] Daniel Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. *ECCV*, 2012.

[8] Zhuoyuan Chen, Hailin Jin, Zhe Lin, Scott Cohen, and Ying Wu. Large displacement optical flow from nearest neighbor fields. *CVPR*, 2013.

[9] Alexey Dosovitskiy, Philipp Fischery, Eddy Ilg, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. *ICCV*, 2015.

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[11] Yoav HaCohen, Eli Shechtman, Dan B Goldman, and Dani Lischinski. Non-rigid dense correspondence with applications for image enhancement. *ACM SIGGRAPH*, 30(70):1–9, 2011.

[12] Berthold Horn and Brian Schunck. Determining optical flow. *Artificial Intelligence*, 16:185–203, 1981.

[13] Simon Korman and Shai Avidan. Coherency sensitive hashing. *ICCV*, 2011.

[14] David G. Lowe. Object recognition from local scale-invariant features. *ICCV*, 1999.

[15] Jiangbo Lu, Hongsheng Yang, Dongbo Min, and Minh Do. Patchmatch filter: Efficient edge-aware filtering meets randomized search for fast correspondence field estimation. *CVPR*, 2013.

[16] Carsten Rother, Vladimir Kolmogorov, Victor Lempitsky, and Martin Szummer. Optimizing binary mrfs via extended roof duality. *CVPR*, 2007.

[17] Deqing Sun, Stefan Roth, and Michael J. Black. Secrets of optical flow estimation and their principles. *CVPR*, 2010.

[18] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. Deepflow: Large displacement optical flow with deep matching. *ICCV*, 2013.

[19] Li Xu, Jiaya Jia, and Yasuyuki Matsushita. Motion detail preserving optical flow estimation. *PAMI*, 16(9):1744–1757, 2012.