Image and Vision Computing 52 (2016) 167–177

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Highly accurate optical flow estimation on superpixel tree ☆

Yinlin Hu [a], Rui Song [a,b], Yunsong Li [a,*], Peng Rao [b], Yangli Wang [a]

[a] State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an, PR China
[b] Key Laboratory of Infrared System Detection and Imaging Technology, Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai, PR China

ARTICLE INFO

Article history:
Received 17 October 2015
Received in revised form 21 May 2016
Accepted 3 June 2016
Available online 20 June 2016

Keywords:
Optical flow
Motion estimation
Cost volume filtering
Super pixel

ABSTRACT

Formulated as a pixel-labeling problem, optical flow estimation using efficient edge-aware filtering has shown great success recently. However, the typical challenge that restricts the range of applicability of this method is the computational complexity mainly caused by the testing of every hypothetical label in the whole label space, which is usually large in an optical flow estimation. In this paper, we present an effective and efficient two-level filter-based optical flow algorithm connected by an accurate non-local matching. With the key observation that the optical flow of the pixels from the same compact superpixels is highly coherent, we propose a novel superpixel tree representation of an image to obtain an accurate superpixel flow. We find that if filtered separately, the candidate label space of the pixels from each superpixel is drastically reduced with the known superpixel flow. We also suggest a refined label selection strategy that is more accurate than the usual winner-takes-all manner. The proposed method, called Highly Accurate flow on Superpixel Tree (HastFlow) is validated on Middlebury and MPI-Sintel, and outperforms all filter-based methods both in accuracy and efficiency.

1. Introduction

Optical flow, also known as a dense motion field, has traditionally been, and continues to be, one of the most fundamental components in many vision tasks, such as segmentation, tracking and surface registration, and object recognition. Since the pioneering work by Horn and Schunck [17] in the early 1980s, many variational methods built on their original work have emerged and many great advances have been recorded recently [23, 27, 33]. However, obtaining a reliable solution remains a challenging problem in situations such as motion boundaries, large displacement motions, and especially in the presence of large occlusion and untextured regions [26].

Contrary to the classic variational methods, an energy minimization method [28] based on Markov random field (MRF) can also elegantly handle the optical flow estimation problem formulated as a pixel-labeling problem. However, the main problem that restricts the range of applicability of this discrete optimization framework is computational complexity, especially when the label space is large, which is common in optical flow estimation.

Recently, partly due to the emerging of efficient local edge-aware filtering [16], the method based on the filtering of the cost volume [25] has shown great success as an efficient alternative to MRF-based global approaches. However, despite the computational complexity of this local filter-based method being independent of the filter kernel size, the user-specified local support region is usually application dependent and obviously only locally optimal. The case of large untextured regions leads to confusing labeling results. To handle the drawback of a window-based filter that is vulnerable to the lack of texture, Yang [34] proposes a novel non-local filter algorithm (TreeFilter) by aggregating cost on a tree structure that has been demonstrated to outperform all the local methods for stereo matching, especially in the case of large untextured regions. Furthermore, by accumulating the distance between the nodes of the constructed tree, the computational complexity of this tree filtering algorithm is extremely low. However, the main problem induced by this tree structure is the less accurate results around the highly-textured regions. This is mainly caused by the fact that pixels around the highly-textured regions receive relatively less support from their neighbors when taking pixels as the tree nodes, as proposed in Ref. [34] (see Fig. 1).

In this study, we propose a two-level filtering framework for optical flow method that combines the non-local matching and local cost aggregation, which hardly impairs the quality but leads to a significant speed-up. The main step we proposed to bridge the gap between non-local and local methods is a minimum spanning tree

☆ This paper has been recommended for acceptance by Etienne Memin.
* Corresponding author.
E-mail addresses: huyinlin@gmail.com (Y. Hu), rsong@xidian.edu.cn (R. Song), ysli@mail.xidian.edu.cn (Y. Li), peng_rao@mail.sitp.ac.cn (P. Rao), yliwang@mail.xidian.edu.cn (Y. Wang).

http://dx.doi.org/10.1016/j.imavis.2016.06.004
0262-8856/

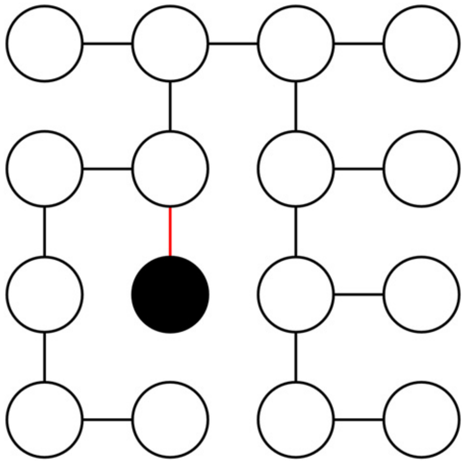

Fig. 1. The typical tree structure from an image using the TreeFilter [34]. The white and black nodes represent some light and dark image pixels, respectively. The single black node is a demonstration of the highly-textured situation. All the edges between the nodes have large support weights, except the red edge between the black node and the corresponding white node due to the poor similarity between them. Furthermore, the filtering result of the black node is vulnerable caused by the less support from its neighbors.

(MST) that takes superpixels [1] as its nodes and is based on the similarity among the superpixels. The motivation of this tree structure is the key observation that the optical flow from the same compact superpixel is highly coherent (as shown in Fig. 2), and the edges between the superpixels in the tree are more robust than the pixel-wise edges. Through an efficient non-local matching procedure on this superpixel-based tree structure, a superpixel flow can be obtained. The superpixel flow can help us to find a rough global optimal initialization and serve as a guide of the local cost aggregation in each superpixel to drastically reduce the candidate label space of each local area.

We make the following contributions: **1)** We propose a more robust tree representation of an image based on super pixels. Compared with the pixel-wise manner, the superpixel tree representation of an image incorporates more robust and edge-adherent spatial scene structure information. **2)** On the basis of this tree structure, we propose a hybrid cost aggregation method for optical flow estimation that combines the advantages of non-local and local methods. Moreover, it improves the robustness over non-local methods contributed by superpixel tree structure, and boosts the efficiency over local methods without quality deterioration under the guidance of superpixel flow. **3)** We also introduce a more accurate label selection strategy that works in conjunction with the winner-takes-all (WTA) manner in the cost volume filtering framework for optical flow estimation. Instead of the WTA label selection scheme, we extend a fitting technique [36] that can significantly reduce the label space without loss of accuracy.

2. Related works

We do not review the entire literature on optical flow and only discuss the related literatures. Benefiting from the significant advances in edge-aware filtering techniques [16, 21, 35], several filter-based methods [25, 29, 37] have recently been proposed for visual correspondence as efficient alternatives to global MRF-based approaches [28]. These effective filtering methods typically aggregate the matching cost over a user-specified local region, but they usually perform less well in large untextured regions. To handle this problem, Yang [34] propose a non-local filtering method based on a tree structure with extremely low computational complexity but is vulnerable in highly-textured regions. Instead of this vulnerable pixel-wise tree structure, a more robust superpixel-based tree structure is proposed in this study.

Many vision applications benefit from superpixels that represents an image as a collection of atomic regions [13, 30, 31]. Closely related to our method, Fernando *et al* [2] solve the optical flow problem in large 3D time-lapse microscopy datasets by defining a Markov random field (MRF) over super-voxels. Chang *et al* [8] and Gkamas *et al* [14] use a superpixel segmentation to guide the dense pixel-based flow estimation, which is less efficient. To handle this problem, Simon *et al*[12] applies a pixel-based technique to the superpixel grid, drastically lowering the resolution that is being processed. In contrast, we tackle the optical flow problem through a cost volume filtering framework based on a superpixel tree structure.

Our superpixel flow is closely related to dense matching. Brox and Malik [6] first introduce a descriptor matching term to the classic variational approach [17] that can better handle large displacements. However, due to the intrinsic problem of the general coarse-to-fine framework, the method usually fails in the case of a tiny structure with motion larger than its own scale. Xu *et al* [33] propose an extended coarse-to-fine refinement framework that integrates matching to refine the flow at each level. To handle the sparsity of descriptor matching, Leordeanu *et al* [19] extend sparse matching to dense matching with locally affine constraint, and Weinzaepfel *et al* [32] integrate a novel deep convolution-matching procedure into the variational algorithm. Here, we present a superpixel-based dense matching procedure using an efficient tree filtering algorithm as mentioned in Ref. [34].

As a core component in image editing and scene correspondence [20], the nearest neighbor field (NNF) has generated much interest recently. The objective of NNF computation is to find one or more nearest neighbors for every patch in a given image against another image. The main challenging in this procedure is the computational complexity. Through the seminal work called PatchMatch [5] and the improving methods [15, 18], the efficiency of computing NNF has advanced remarkably. With a different objective, the computed NNF is different from the motion field, especially in the aspects of edge preservation and spatial smoothness. However, Chen *et al* [9] use the computed NNF as a hint for motion segmentation before variational optimization and achieve excellent results on Middlebury [3]. Bao *et al* [4] adapt the PatchMatch algorithm to optical flow using an edge-preserving patch-matching measurement. Our method is closer to the approach of [22], which proposes a filtering framework based on a randomized search paradigm and propagation from neighboring superpixels inspired by PatchMatch. Unlike the previous work, we use a guide from the superpixel flow to reduce the local search label space rather than randomized search, which is more suitable for parallelization.

3. Problem setup

In this section, we briefly present the problem formulation and summarize the work of Rhemann *et al*[25], who propose the cost volume filtering framework for visual correspondence, and the work of Yang [34], who propose a novel non-local filtering scheme.

Given a pair of images I_t and I_{t+1}, the optical flow estimation is formulated as a general labeling problem with the goal of assigning each pixel i with coordinates (x_i, y_i) to a label l from the label set $\mathcal{L} = \{1, \ldots, L\}$. $|\mathcal{L}| = L$ is the label space size. For the optical flow, the label l corresponds to vectors (u, v), where u and v correspond to the displacement in the x and y directions, respectively.

This filter-based framework consists of three steps: 1) cost volume construction, 2) cost aggregation, and 3) label selection. The cost volume C is a 3D array that consists of L cost slices that store the costs of every pixel $i = (x_i, y_i)$ for choosing a hypothetical label