Tal Hassner · Ce Liu *Editors*

Dense Image Correspondences for Computer Vision

Springer

Tal Hassner • Ce Liu
Editors

Dense Image Correspondences for Computer Vision

Springer

Editors
Tal Hassner
Department of Mathematics
and Computer Science
The Open University of Israel
Raanana, Israel

Ce Liu
Google Research
Cambridge, MA, USA

ISBN 978-3-319-23047-4 ISBN 978-3-319-23048-1 (eBook)
DOI 10.1007/978-3-319-23048-1

Library of Congress Control Number: 2015953102

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

To my wife, Osnat, and my children, Ben, Ella, Daniel, and May
With love beyond words
– T.H.

To my wife, Irene, and my daughter, Mary
– C.L.

Preface

Correspondence estimation is a task of matching pixels of one image with those of another. When referring to *dense* correspondence estimation, the emphasis is on finding suitable matches (*correspondences*) for every one of those pixels.

Throughout much of the history of computer vision as a field of research, work on dense correspondence estimation has mostly been motivated by two specific problems: stereo vision, in which the pixels in one view of a 3D scene are matched with pixels in another view of the same scene to determine displacements and reason about 3D structure; and optical flow, in which the two images are taken from the same camera, but at different points in time. Many solutions have been offered for both these tasks, varying in their algorithmic design or the assumptions they make on the nature of the scene and imaging conditions. Implicit in most, however, is the *same scene* assumption: that is, that the two images involved capture the same physical scene, with possible differences due to independent scene motion.

This assumption plays an important role in defining the criteria for how pixels should be matched. If the same scene is visible in both images, then any physical scene point is expected to appear the same in both views. This similar appearance can therefore be used to match the pixels which capture its appearance in both images. Classical optical flow methods often make this concrete by using the *brightness constancy* assumption which interprets similar appearance as similar local patterns of pixel intensities.

In recent years, however, there has been a growing interest in breaking away from this same scene assumption and designing methods for correspondence estimation even in cases where the two images capture entirely different scenes. The rational for doing so is not entirely obvious and requires some explanation. For one thing, matching pixels between images of different scenes implies a harder problem: If the two images present different scenes, a physical point appearing in one image will obviously not appear in the other image and in particular cannot be expected to appear the same in both images. Therefore, the brightness constancy assumption cannot be applied in such cases and new criteria must be established in order to determine when two pixels actually match.

But even if correspondences could reliably be established even under these challenging circumstances, there is still the question of why this would even make sense. It is not immediately obvious what disparities can say about scene structure when two scenes are involved. Similarly, both scene and camera motion are not necessarily meaningful in such cases.

This book is intended to present the problem, solutions and applications of dense correspondence estimation with a particular emphasis on cross scene correspondences. Its chapters tackle the two issues raised above: first, by describing techniques designed to enable correspondence estimation under increasingly challenging conditions and second, by showing what can be done with these correspondences. The book is accordingly divided into the following two parts.

In Part I, we focus on *how* dense correspondences may be estimated. Chapter "Introduction to Dense Optical Flow" provides a survey of classical optical flow methods with an emphasis on the pioneering work of Horn/Schunck. Chapter "SIFT Flow: Dense Correspondence Across Scenes and Its Applications" takes dense correspondences beyond the same scene settings, by introducing the *SIFT flow* method. In chapter "Dense, Scale-Less Descriptors", our focus shifts from the correspondence estimation method to the per pixel representation. It presents the *Scale Less SIFT* descriptor which widens the application of SIFT flow to images presenting information at different scales. A different approach to scale invariance in cross scene correspondence estimation, the *Scale-Space SIFT flow*, is described in chapter "Scale-Space SIFT Flow". A descriptor design approach intended to improve the discriminative quality of the per pixel representation is provided in chapter "Dense Segmentation-Aware Descriptors", which describes *segmentation-aware* descriptors. Chapter "SIFTpack: A Compact Representation for Efficient SIFT Matching" takes on a different concern: the storage and computational requirements often involved in dense correspondence estimation. The *SIFTpack* representation offers a compact and more efficient alternative to the Dense SIFT representation used by SIFT flow and related methods. Finally, whereas methods such as SIFT flow use a graph-based search for alignment, chapter "In Defense of Gradient-Based Alignment on Densely Sampled Sparse Features" explores an alternative approach of *gradient-based alignment* by continuous optimization.

In Part II, we focus on *why* dense correspondences are useful, even when they are computed between images of different scenes, by showing how they may be used to solve a wide range of computer vision problems. Specifically, chapter "From Images to Depths and Back" looks back to one of the early uses of cross scene dense correspondence estimation for estimating scene depth from a single view. The more recent *Depth Transfer* method for single view depth estimation by dense correspondence estimation is presented in chapter "DepthTransfer: Depth Extraction from Video Using Non-parametric Sampling". Single image scene parsing by the *Label Transfer* method is described in chapter "Nonparametric Scene Parsing via Label Transfer". The Label Transfer approach assumes many reference images with matching label information, which is transferred through dense correspondences to novel query images. The *Joint Inference* approach described in chapter "Joint Inference in Image Datasets via Dense Correspondence" shows how

this approach can be applied even when labels are available for only a few reference images. Finally, chapter "Dense Correspondences and Ancient Texts" takes dense correspondence estimation to an entirely different imaging domain and shows how dense correspondences may be used to process challenging ancient, handwritten texts.

Taken as a whole, this book shows that accurate dense correspondence estimation is possible, even under challenging settings, and can be key to solving image understanding problems in problem domains far beyond stereo and optical flow. We hope that this book will make the methods and applications of dense correspondence estimation accessible. Going beyond existing work, we would like to see this book motivate the development of new, more accurate, more robust and more efficient dense correspondence estimation techniques.

The editors are most grateful to all the friends and colleagues who have supported this book by contributing their work for its chapters: Xiang Bai, Ronen Basri, Hilton Bristow, Nachum Dershowitz, Alexandra Gilinsky, Sing Bing Kang, Kevin Karsch, Iasonas Kokkinos, Simon Lucey, Viki Mayzels, Francesc Moreno-Noguer, Weichao Qiu, Miki Rubinstein, Gil Sadeh, Alberto Sanfeliu, Daniel Stökl Ben-Ezra, Antonio Torralba, Eduard Trulls, Zhuowen Tu, Xinggang Wang, Lior Wolf, Jenny Yuen, Alan Yuille and Lihi Zelnik-Manor.

Tal Hassner thanks the Open University of Israel (OUI) and, in particular, the chief of its research authority, Daphna Idelson, for their generous support for this book. He also gratefully acknowledges the longtime guidance, support and friendship of Ronen Basri, Michal Irani, Lior Wolf and Lihi Zelnik-Manor.

Ce Liu thanks Harry Shum, Rick Szeliksi, Bill Freeman, Ted Adelson, Antonio Torralba and Yair Weiss for their advice and support in his life. He is grateful to his collaborators, Antonio Torralba, Jenny Yuen, Miki Rubinstein, Marshall Tappen, Kevin Karsch, Jaechul Kim and Philip Isola, on the topic of dense correspondences.

Raanana, Israel — Tal Hassner
Cambridge, MA, USA — Ce Liu

Contents

Part I
Establishing Dense Correspondences

Introduction to Dense Optical Flow

Ce Liu

Abstract Before the notion of motion is generalized to arbitrary images, we first give a brief introduction to motion analysis for videos. We will review how motion is estimated when the underlying motion is *slow* and *smooth*, especially the Horn–Schunck (Artif Intell 17:185–203, 1981) formulation with robust functions. We show step-by-step how to optimize the optical flow objective function using iteratively reweighted least squares (IRLS), which is equivalent to conventional Euler–Lagrange variational approach but more succinct to derive. Then we will briefly discuss how motion is estimated when the slow and smooth assumption becomes invalid, especially how large displacement motion is estimated.

1 Introduction

Motion estimation is one of the corner stones of computer vision. It is widely used in video processing to compress videos and to enhance video qualities, and also used in 3D reconstruction, object/event tracking, segmentation, and recognition.

Although video cameras are able to record pixels of the moving objects, motion is unfortunately not recorded directly. Although the amount of motion can be physically measured at very high accuracy in a lab setup, motion remains as a percept instead of direct measurement for general videos. The challenge of motion estimation is therefore to obtain motion that is consistent with human perception.

Although multiple representations of motion have been invented, the most popular ones are *parametric motion* such as affine and homography (projective) where the displacement of pixels undergoes certain parametric forms, or *optical flow fields*, where every pixel has its own displacement vector. These two representations mainly differ in how the motion fields are regularized across the image lattice, while the optimization and initialization strategies are almost the same. Therefore, in this chapter we focus on optical flow estimation. For parametric motion estimation, please refer to [1].

C. Liu (✉)
Google Research, Cambridge, MA, USA
e-mail: celiu@google.com

T. Hassner, C. Liu (eds.), *Dense Image Correspondences for Computer Vision*, DOI 10.1007/978-3-319-23048-1_1

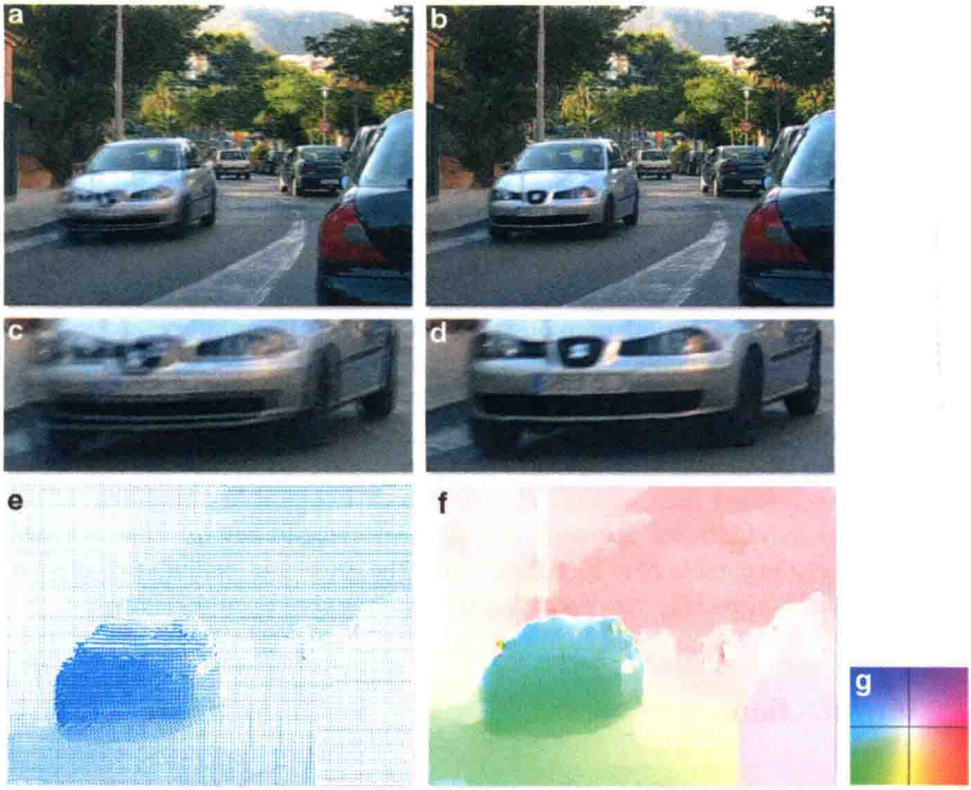

Fig. 1 Illustration of optical flow fields. (**a**) Superimposition of two input frames. (**b**) Superimposition of frame 1 and warped frame 2 according to the estimated flow field. (**c**) and (**d**) are the zoomed-in versions of (**a**) and (**b**), respectively. Notice the double imaging in (**a**) and (**c**) due to the motion between the two frames and the sharp boundaries in (**b**) and (**d**) as the motion is canceled via warping. This is a common way to inspect motion on printed paper, while flipping back and forth two frames is a better way to inspect motion on a digital display. (**e**) and (**f**) are two different ways to visualize the flow fields. In (**e**) flow fields are plotted as vectors on sparse grid. Consequently, it is challenging to visualize a dense flow field with large magnitude. In (**f**) flow fields are visualized via a color-coding scheme in (**g**) [2], where *hue* indicates orientation and *saturation* indicates magnitude. It has become the standard of optical flow visualization since it is possible to see the motion of every pixel. The two frames are from the MIT ground-truth motion database [15]

The effect of motion is illustrated in Fig. 1. For two adjacent frames in a video sequence, the correct flow field should be able to cancel the underlying motion such that the second frame should be identical to the first frame after being warped according to the flow field. In addition, the discontinuity of the flow field should reflect the boundary of the objects in the scene. These are general guidelines to inspect the correctness of the estimated flow field when the ground-truth flow field is absent.

In this chapter we will introduce the optical flow formulation from the basic brightness constancy and smoothness assumption and derive the optical flow estimation algorithm via *incrementalization* and *linearization* (Taylor expansion) of

the objective function. The optimization is derived using iteratively reweighted least squares (IRLS), which is equivalent to the conventional Euler–Lagrange variational approach but is more succinct to derive. Initialization is a key component of optical flow estimation that has been often overlooked. We will discuss about the conventional coarse-to-fine scheme and recent advances using feature matching for initialization to account for fast moving scenes.

For the readers who are interested in knowing more about the literature of optical flow estimation, there are a number of references. An early survey and evaluation of optical flow algorithms can be found in [4]. Various forms of motion estimation are discussed in Rick Szeliski's book [20], while details of optical flow optimization strategies are discussed in Ce Liu's PhD Thesis [14].

2 Slow and Smooth Optical Flow

The motion of the moving scenes both in the nature and in the man-made world takes drastically different forms. Most of the things in our surroundings do not move typically, such as the land, buildings, walls, and desks. The motion of vehicles can be regarded as moving planes from side view except for the wheels. Animals and humans' motion is more complicated with degrees of articulation. The motion of water, fire, and explosion can be so complicated that it is beyond human comprehension. To make the motion estimation problem tractable, a common assumption is that motion is *slow* and *smooth*, namely the pixels tend to stay still or move at low speed, and neighbor pixels tend to move together.

2.1 Basic Formulation

Let image lattice be $p = (x, y) \in \Lambda$, and the two images to match be $I_1(p)$ and $I_2(p)$. Denote $w_p = (u_p, v_p)$ the flow vector at pixel p, where u_p and v_p are the horizontal and vertical components of the flow fields, respectively.

As motion before, the optical flow field should be able to align the two images along with the flow field. The objective function of optical flow is [12]

$$E(w) = \sum_p \psi(|I_1(p) - I_2(p + w_p)|^2) + \lambda \sum_p \phi(|\nabla u_p|^2 + |\nabla v_p|^2), \tag{1}$$

where the first term is often called the *data term* and the second term is called the *smoothness term*. λ is a coefficient that balances the two terms. In this equation, $\psi(\cdot)$ and $\phi(\cdot)$ are both robust functions [6], which can take following forms:

- L2 norm: $\psi(z^2) = z^2$
- L1 norm: $\psi(z^2) = \sqrt{z^2 + \varepsilon^2}$
- Lorentzian: $\psi(z^2) = \log(1 + \gamma z^2)$

The same functions can be chosen for the smoothness robust function ϕ as well. Both L1 and Lorentzian forms make the function robust, namely to be able to account for matching outliers in the data term and to encourage piecewise smooth discontinuities. This is often called L1 total variation optical flow [9].

It is worth noting that the two images I_1 and I_2 are not limited to grayscale images. If they can be multiple-channel images (such as RGB, HSV, and YUV), an extra summation over the channels is needed in the data term. The images may also contain some image features such as gradients and other features from linear or nonlinear filtering. When gradients are used, then gradient constancy is implied, which can make the flow more stable at the presence of global illumination change.

This objective function in Eq. (1) is very difficult to minimize because of the warping function $I_2(p + w_p)$. To deal with warping, we follow the typical incrementation and linearization strategy. First, the objective function can be rewritten to optimize over incremental $dw = (du, dv)$ of the flow field:

$$E(w, dw) = \sum_p \psi(|I_1(p) - I_2(p + w_p + dw_p)|^2) + \lambda \sum_p \phi(|\nabla(u_p + du_p)|^2 + |\nabla(v_p + dv_p)|^2). \tag{2}$$

Second, we can linearize the warping using Taylor expansion

$$I_2(p + w_p + dw_p) - I_1(p) \approx I_t(p) + I_x(p)du_p + I_y(p)dv_p, \tag{3}$$

where

$$I_t(p) = I_2(p + w_p) - I_1(p), \tag{4}$$

$$I_x(p) = \frac{\partial}{\partial x} I_2(p + w_p), \tag{5}$$

$$I_y(p) = \frac{\partial}{\partial y} I_2(p + w_p). \tag{6}$$

Now the objective function becomes

$$E(w, du, dv) = \sum_p \psi(|I_t(p) + I_x(p)du_p + I_y(p)dv_p|^2) + \lambda \sum_p \phi(|\nabla(u + du)_p|^2 + |\nabla(v + dv)_p|^2). \tag{7}$$

Our goal is to rewrite Eq. (7) in vector and matrix forms. We define the robust function Ψ and Φ as applying the robust function to each element of a vector:

$$\Psi(X) = [\psi(X_1), \psi(X_2), \ldots, \psi(X_n)]^T, \tag{8}$$

$$\Phi(X) = [\phi(X_1), \phi(X_2), \ldots, \phi(X_n)]^T. \tag{9}$$

To be succinct, we denote XY and X^2 as element-wise multiplication and element-wise square for vectors

$$XY = [X_1Y_1, X_2Y_2, \ldots, X_nY_n]^T, \tag{10}$$

$$X^2 = [X_1^2, X_2^2, \ldots, X_n^2]^T. \tag{11}$$

In this way Eq. (7) can be rewritten in a vector form

$$\begin{aligned} E(w, du, dv) = {} & \mathbf{1}^T \Psi\Big((I_t + I_x du + I_y dv)^2\Big) \\ & + \lambda \mathbf{1}^T \Phi\Big(\big(\mathbf{D}_x(u + du)\big)^2 + \big(\mathbf{D}_y(u + du)\big)^2 \\ & + \big(\mathbf{D}_x(v + dv)\big)^2 + \big(\mathbf{D}_y(v + dv)\big)^2\Big), \end{aligned} \tag{12}$$

where $\mathbf{D}_x$ and $\mathbf{D}_y$ are matrices corresponding to x- and y-derivative filters, such as $[-1, 1]$ filters.

2.2 Optimization via Iteratively Reweighted Least Squares

The typical approach that can be found in the optical flow literature is to use Euler–Lagrange to find fixed point of the partial differential equations (PDEs). We are going to show how to derive using Euler–Lagrange in the next subsection. Here, we are going to use IRLS to directly optimize the objective function in Eq. (7).

To simplify the notations, denote

$$\begin{aligned} \Psi' &= \Psi'\Big((I_t + I_x du + I_y dv)^2\Big), \\ \boldsymbol{\Psi}'_{xx} &= \mathrm{diag}(\Psi' I_x I_x),\ \boldsymbol{\Psi}'_{xy} = \mathrm{diag}(\Psi' I_x I_y),\ \boldsymbol{\Psi}'_{yy} = \mathrm{diag}(\Psi' I_y I_y), \\ \boldsymbol{\Psi}'_{xt} &= \mathrm{diag}(\Psi' I_x I_t),\ \boldsymbol{\Psi}'_{yt} = \mathrm{diag}(\Psi' I_y I_t), \\ \boldsymbol{\Phi}' &= \mathrm{diag}\Big(\Phi\Big(\big(\mathbf{D}_x(u + du)\big)^2 + \big(\mathbf{D}_y(u + du)\big)^2 \\ &\qquad + \big(\mathbf{D}_x(v + dv)\big)^2 + \big(\mathbf{D}_y(v + dv)\big)^2\Big)\Big), \\ \mathbf{L} &= \mathbf{D}_x^T \boldsymbol{\Phi}' \mathbf{D}_x + \mathbf{D}_y^T \boldsymbol{\Phi}' \mathbf{D}_y. \end{aligned} \tag{13}$$

Taking the derivative of the objective function w.r.t. du and setting it to be zero, we obtain