

A Quantitative Analysis of Current Practices in Optical Flow Estimation and the Principles Behind Them

Deqing Sun · Stefan Roth · Michael J. Black

Received: 10 January 2013 / Accepted: 10 July 2013
© The Author(s) 2013. This article is published with open access at Springerlink.com

Abstract The accuracy of optical flow estimation algorithms has been improving steadily as evidenced by results on the Middlebury optical flow benchmark. The typical formulation, however, has changed little since the work of Horn and Schunck. We attempt to uncover what has made recent advances possible through a thorough analysis of how the objective function, the optimization method, and modern implementation practices influence accuracy. We discover that “classical” flow formulations perform surprisingly well when combined with modern optimization and implementation techniques. One key implementation detail is the median filtering of intermediate flow fields during optimization. While this improves the robustness of classical methods it actually leads to higher energy solutions, meaning that these methods are not optimizing the original objective function. To understand the principles behind this phenomenon, we derive a new objective function that formalizes the median filtering heuristic. This objective function includes a non-local smoothness term that robustly integrates flow estimates over large spatial neighborhoods. By modifying this

new term to include information about flow and image boundaries we develop a method that can better preserve motion details. To take advantage of the trend towards video in wide-screen format, we further introduce an asymmetric pyramid downsampling scheme that enables the estimation of longer range horizontal motions. The methods are evaluated on the Middlebury, MPI Sintel, and KITTI datasets using the same parameter settings.

Keywords Optical flow estimation · Practices · Median filtering · Non-local term · Motion boundary

1 Introduction

The field of optical flow estimation is making steady progress as evidenced by the increasing accuracy of current methods on the Middlebury optical flow benchmark (Baker et al. 2007). After over 30 years of research, these methods have obtained an impressive level of reliability and accuracy (Wedel et al. 2008b, 2009; Werlberger et al. 2009; Xu et al. 2012; Zimmer et al. 2009). *But what has led to this progress?* The majority of today’s methods strongly resemble the original formulation of Horn and Schunck (HS, 1981). They combine a data term that assumes constancy of some image property with a spatial term that models how the flow is expected to vary across the image. An objective function combining these two terms is then optimized. Given that this basic structure is unchanged since HS, what has enabled the performance gains of modern approaches?

The paper has three parts. In the first, we perform a study of recent optical flow methods and models. The most accurate methods on the Middlebury flow dataset make different choices about how to model the objective function, how to approximate this model to make it computationally tractable,

Portions of this work were performed when DS and MJB were at Brown University.

D. Sun (✉)
School of Engineering and Applied Sciences, Harvard University,
Cambridge, MA, USA
e-mail: dqsun@seas.harvard.edu

S. Roth
Department of Computer Science, TU Darmstadt,
Darmstadt, Germany
e-mail: sroth@cs.tu-darmstadt.de

M. J. Black
Max Planck Institute for Intelligent Systems,
Tübingen, Germany
e-mail: black@is.mpg.de

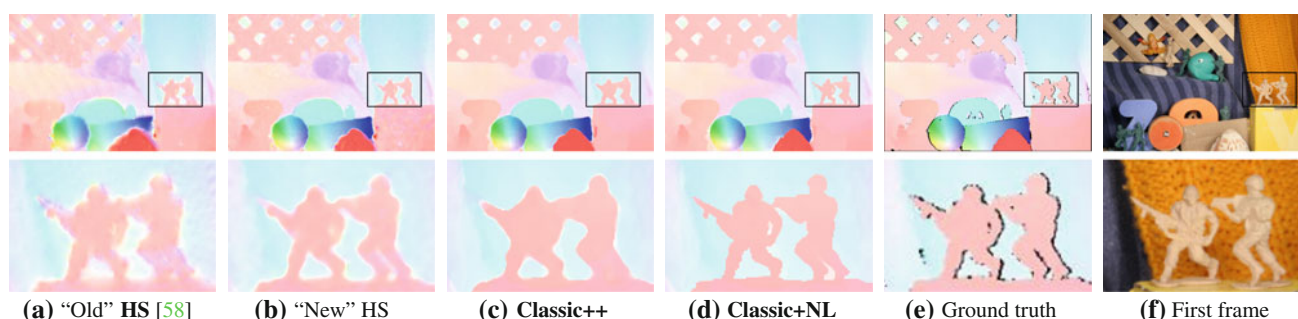


Fig. 1 Estimated optical flow on the Middlebury test “Army” sequence. *Left to right*: **a** an old implementation of the Horn and Schunck (**HS**) method (Sun et al. 2008), **b** a new implementation with current practices, **c** a modern implementation of a robust version, **d** an improved model that uses a non-local spatial term to robustly inte-

grate information over a large spatial neighborhood, **e** ground truth from the Middlebury website (downsampled and JPEG compressed; original ground truth is withheld), and **f** the first frame. Color coding as in (Baker et al. 2007), shown in Fig. 4c. Average end-point error (EPE): **a** 0.22, **b** 0.12, **c** 0.09, and **d** 0.08

and how to optimize it. Since most published methods change *all* of these properties at once, it can be difficult to know which choices are most important. To address this, we define a baseline algorithm that is “classical”, in that it is a direct descendant of the original **HS** formulation, and then systematically vary the model and method using different techniques from the art. The results are surprising. We find that only a small number of key choices produce statistically significant improvements and that they can be combined into a very simple method that achieves reasonable accuracy. More importantly, our analysis reveals what makes current flow methods work so well.

Part two examines the *principles* behind this success. We find that one algorithmic choice produces the most significant improvements: applying a median filter to intermediate flow values during incremental estimation and warping (Wedel et al. 2008b, 2009). While this heuristic improves the accuracy of the recovered flow fields, it actually *increases* the energy of the objective function. This suggests that what is being optimized is actually a new and different objective. Using observations about median filtering and L1 energy minimization from Li and Osher (2009), we formulate a new *non-local term* that is added to the original, classical objective. This new term goes beyond standard local (pairwise) smoothness to robustly integrate information over large spatial neighborhoods. We show that minimizing this new energy approximates the original optimization with the heuristic median filtering step. Note, however, that the new objective falls outside our definition of classical methods.

Once the median filtering heuristic is formulated as a non-local term in the objective, we immediately recognize how to modify and improve it. In part three we show how information about image structure and flow boundaries can be incorporated into a weighted version of the non-local term to prevent over-smoothing across boundaries. By incorporating structure from the image, this weighted version does not

suffer from some of the errors produced by median filtering and better preserves motion boundaries. Figure 1 illustrates optical flow estimates for a range of methods from a “basic” **HS** method to our proposed **Classic+NL** method.

Finally we observe that the classical methods all go beyond the original **HS** algorithm by using a spatial pyramid to cope with large motions. The classical pyramid downsamples the image equally in both the horizontal and vertical direction, typically until some minimum image dimension is reached. With today’s wide-aspect ratio video, we point out that an asymmetric approach can be employed resulting in a pyramid that downsamples more in the horizontal direction than in the vertical one. This effectively allows the estimation of larger horizontal motions. This simple change results in significant improvements on the wide-aspect-ratio video in the KITTI (Geiger et al. 2012) and MPI Sintel (Butler et al. 2012) datasets.

At the time of writing our previous conference paper (Sun et al. 2010a, March), the resulting approach was ranked 1st in both angular and end-point errors in the Middlebury evaluation. At the writing of this paper (Sep. 2012), the method, **Classic+NL**, ranks 13th in both AAE and EPE. Several recent and high-ranking methods directly build on **Classic+NL**, such as layered models (Sun et al. 2010b, 2012, 2013), methods with more advanced motion prior models (Chen et al. 2012; Jia et al. 2011), efficient optimization schemes for the non-local term (Krähenbühl and Koltun 2012), and better initialization to deal with large displacement optical flow (Chen et al. 2013).

Compared to the conference version (Sun et al. 2010a), this paper includes many more detailed results and analyses. In addition to an expanded literature review we compare our proposed method to the closely related non-local total variation method (Werlberger et al. 2010). We discuss the limitations of our method in dealing with occlusions and fast moving objects. We report results on the MIT HAMA

data set (Liu et al. 2008) and find that the results are consistent with those on Middlebury. We also test our methods on the MPI Sintel (Butler et al. 2012) and KITTI (Geiger et al. 2012) datasets, which offer greater challenges. Using the same parameters tuned on the Middlebury training set, our method performs well on these new datasets, particularly using an asymmetric pyramid.

In summary, the contributions of this paper are to (1) analyze current flow models and methods to understand which design choices matter; (2) formulate and compare several classical objectives descended from **HS** using modern methods; (3) formalize one of the key heuristics and derive a new objective function that includes a non-local spatial smoothness term; (4) modify this new objective to produce a competitive method; (5) extend spatial pyramids to exploit the extra width of high-definition and letterbox videos. In doing so, we provide a “recipe” for others studying optical flow that can guide their design choices. Finally, to enable comparison and further innovation, we provide a public MATLAB implementation (<http://www.cs.brown.edu/people/dqsun>; last accessed 24 July 2013).

2 Previous Work

It is important to separately analyze the contributions of the objective function that defines the problem (Sect. 2.1) and the optimization algorithm and implementation used to minimize it (Sect. 2.2). The **HS** formulation, for example, has long been thought to be highly inaccurate. Barron et al. (1994) reported an average angular error (AAE) of $\sim 30^\circ$ on the “Yosemite” sequence. This confounds the objective function with the particular optimization method proposed by Horn and Schunck. Horn and Schunck noted that the correct way to optimize their objective is by solving a system of linear equations as is common today. This was impractical on the computers of the day, hence they used a heuristic method. In fact, Barron et al. note that the original **HS** derivatives were implemented crudely and report a modified version of **HS** with AAE around 11° . When optimized with today’s methods, the **HS** objective achieves surprisingly competitive results (Geiger et al. 2012) despite the expected over-smoothing and sensitivity to outliers. The reported accuracy of a method is jointly determined by the objective function, the optimization techniques, the implementation details, and the parameter tuning/learning (cf. Marr 1982; Szeliski 2010). We review related research in the context of the first three aspects below.

2.1 Models

The global formulation of optical flow introduced by Horn and Schunck (1981) relies on both brightness constancy and

spatial smoothness assumptions, but suffers from the fact that their quadratic formulation is not robust to outliers. Shulman and Herve (1989) use an L1 penalty instead to preserve flow discontinuities. Black and Anandan (1996) introduce a robust framework to deal with outliers in both the data and the spatial terms. Subsequently, many different robust functions have been explored (Brox et al. 2004; Lempitsky et al. 2008; Sun et al. 2008) and it remains unclear which is best. We refer to all these spatially-discrete formulations derived from **HS** as “classical.” We systematically explore variations in the formulation and optimization of these approaches. The surprise is that the classical model, appropriately implemented, remains fairly competitive.

There are many formulations beyond the classical ones that we do not consider here. Significant ones use oriented smoothness (Nagel and Enkelmann 1986; Sun et al. 2008; Wedel et al. 2009; Zimmer et al. 2011, 2009), rigidity constraints (Wedel et al. 2008a, 2009), an over-parameterized smoothness term (Nir et al. 2008), or image segmentation (Black and Jepson 1996; Lei and Yang 2009; Xu et al. 2008; Zitnick et al. 2005). While they deserve similar careful consideration, we expect many of our conclusions to carry forward. Note that one can select among a set of models or methods for a given sequence (Mac Aodha et al. 2010), instead of finding a “best” model for all the sequences.

2.2 Methods

Many of the implementation details that are thought to be important date back to the early days of optical flow. Current best practices include coarse-to-fine estimation to deal with large motions (Bergen et al. 1992; Brox et al. 2004), texture decomposition (Wedel et al. 2008a,b) or high-order filter constancy (Adelson et al. 1984; Brox et al. 2004; Glaer et al. 1983; Lempitsky et al. 2010; Zimmer et al. 2009) to reduce the influence of lighting changes, incremental warping (Bergen et al. 1992), warping with bicubic interpolation (Lempitsky et al. 2008; Wedel et al. 2008b), temporal averaging of image derivatives (Horn 1986; Wedel et al. 2008b), graduated non-convexity (Blake and Zisserman 1987) to minimize non-convex energies (Black and Anandan 1996; Sun et al. 2008), and median filtering after each incremental estimation step to remove outliers (Wedel et al. 2008b).

This median filtering heuristic is of particular interest as it makes non-robust methods more robust and improves the accuracy of all methods we tested. The effect on the objective function and the underlying reason for its success have not previously been analyzed. Least median squares estimation can be used to robustly reject outliers in flow estimation (Bab-Hadiashar and Suter 1998), but previous work has focused on the data term.

Related to median filtering, and our new non-local term, is the use of bilateral filtering to prevent smoothing across

motion boundaries (Xiao et al. 2006). This approach separates a variational method into two filtering update stages, and replaces the original anisotropic diffusion process with multi-cue driven bilateral filtering. As with median filtering, the bilateral filtering step changes the original energy function.

Models that are formulated with an L1 robust penalty are often coupled with specialized total variation (TV) optimization methods (Zach et al. 2007). Here we focus on generic optimization methods that can apply to most models and find that the estimated flow fields are as accurate as the reported results for specialized methods.

Despite recent algorithmic advances, there is a lack of publicly available, easy to use, and accurate flow estimation software. The GPU4Vision project (<http://gpu4vision.icg.tugraz.at>; last accessed 24 July 2013) has made a substantial effort to change this and provides executable files for several accurate methods (Wedel et al. 2008a,b, 2009; Werlberger et al. 2009). The dependence on the GPU and the lack of source code are limitations. Since the publication of our conference paper, our public MATLAB code has been used by both researchers to develop new optical flow algorithms (Adato et al. 2011; Chen et al. 2012, 2013; Jia et al. 2011; Krähenbühl and Koltun 2012) and practitioners to use optical flow for different applications (Humayun et al. 2011; Lin and Fisher 2012; Niu et al. 2012). Currently other available optical-flow software includes (<http://lmb.informatik.uni-freiburg.de/resources/software.php>; last accessed 24 July 2013 <http://people.csail.mit.edu/celiu/OpticalFlow/>; last accessed 24 July 2013 <http://www.cse.cuhk.edu.hk/leojia/projects/flow/>; last accessed 24 July 2013).

3 Classical Models

As is common to “classical” methods we only address the two-frame optical flow estimation problem. We write the classical optical flow objective function in its spatially discrete form as

$$E(\mathbf{u}, \mathbf{v}) = \sum_{i,j} \left\{ \rho_D(I_1(i, j) - I_2(i + u_{i,j}, j + v_{i,j})) + \lambda [\rho_S(u_{i,j} - u_{i+1,j}) + \rho_S(u_{i,j} - u_{i,j+1}) + \rho_S(v_{i,j} - v_{i+1,j}) + \rho_S(v_{i,j} - v_{i,j+1})] \right\}, \quad (1)$$

where \mathbf{u} and \mathbf{v} are the horizontal and vertical components of the optical flow field to be estimated from images I_1 and I_2 , i, j indexes a particular image pixel location, $u_{i,j}$ and $v_{i,j}$ are elements of \mathbf{u} and \mathbf{v} respectively, λ is a regularization parameter, and ρ_D and ρ_S are the data and spatial penalty functions. We consider three different penalty functions: (1) the quadratic HS penalty $\rho(x) = x^2$; (2) the Charbonnier penalty $\rho(x) = \sqrt{x^2 + \epsilon^2}$ (Bruhn et al. 2005), a differentiable variant of the absolute value, the most robust con-

vex function; and (3) the Lorentzian $\rho(x) = \log(1 + \frac{x^2}{2\sigma^2})$, which is a non-convex robust penalty used by Black and Anandan (1996). We refer to the robust formulation with the Lorentzian penalty as **BA** (short for Black and Anandan). Note that this classical model is related to a standard pairwise Markov random field (MRF) based on a 4-neighborhood (Geman and Geman 1984).

In the remainder of this section we define a baseline method using several techniques from the literature. This is not the “best” method, but includes modern techniques and will be used for comparison. We only briefly describe the main choices, which are explored in more detail in the following section and the cited references.

Quantitative results are presented throughout the remainder of the text. In all cases we report the average end-point error (EPE) on the Middlebury training and test sets, depending on the experiment.

3.1 Baseline Methods

To gain robustness against lighting changes, we follow Wedel et al. (2008b) and apply the Rudin–Osher–Fatemi (ROF; Rudin et al. 1992) structure texture decomposition method to pre-process the input sequences and linearly combine the texture and structure components (in the proportion 20:1). The parameters are set according to Wedel et al. (2008b).

Optimization is performed using a standard incremental multi-resolution technique (e.g., Black and Anandan 1996; Brox et al. 2004) to estimate flow fields with large displacements. The optical flow estimated at a coarse level is used to warp the second image toward the first at the next finer level, and a flow increment is calculated between the first image and the warped second image. The standard deviation of the Gaussian anti-aliasing filter is set to be $\frac{1}{\sqrt{2d}}$, where d denotes the downsampling factor. Each level is recursively downsampled from its nearest lower level. In building the pyramid, the downsampling factor is not critical as pointed out in the next section; here we use the settings of Sun et al. (2008), which uses a factor of 0.8 in the final stages of the optimization. For the basic pyramid scheme, we adaptively determine the number of pyramid levels so that the top level has a width or height of around 20–30 pixels. At each pyramid level, we perform 10 warping steps to compute the flow increment.

At each warping step, we linearize the data term once, which involves computing terms of the type $\frac{\partial}{\partial x} I_2(i + u_{i,j}^k, j + v_{i,j}^k)$, where $\partial/\partial x$ denotes the partial derivative in the horizontal direction, u^k and v^k denote the current flow estimate at iteration k . As suggested by Wedel et al. (2008b), we compute the derivatives of the second image using the 5-point derivative filter $\frac{1}{12}[-1 \ 8 \ 0 \ -8 \ 1]$, and warp the second image and its derivatives toward the first using the current flow estimate

by bicubic interpolation. We then compute the spatial derivatives of the first image, compute the average of these and the corresponding warped derivatives of the second image (cf. [Álvarez et al. 2007](#); [Horn 1986](#)), and use these in place of $\frac{\partial I_2}{\partial x}$. For pixels moving out of the image boundaries, we set both their corresponding temporal and spatial derivatives to zero. After each warping step, the flow update is computed, and then we apply a 5×5 median filter to the newly computed flow field to remove outliers ([Wedel et al. 2008b](#)).

For the Charbonnier (**Classic-C**) and Lorentzian (**Classic-L**) penalty function, we use a graduated non-convexity scheme (GNC; [Blake and Zisserman 1987](#)) as described by [Sun et al. \(2008\)](#). First, we replace the robust penalty functions by quadratic penalty functions and obtain a quadratic formulation of the objective function, $E_Q(\mathbf{u}, \mathbf{v})$. Then we linearly combine the quadratic penalty function with the desired robust penalty function and gradually change the weighting of the two terms to reach the desired robust penalty function. In practice, we use a three-stage GNC scheme, with the objective functions for the first, second, and third stages being $E_Q(\mathbf{u}, \mathbf{v})$, $\frac{1}{2}(E_Q(\mathbf{u}, \mathbf{v}) + E(\mathbf{u}, \mathbf{v}))$, and $E(\mathbf{u}, \mathbf{v})$ respectively. The output of a previous stage serves as the initialization to the next stage. The standard deviations of the corresponding quadratic penalty function are set to be 1 for the Charbonnier penalty and, for the Lorentzian, are taken to be the same as the σ value used in the Lorentzian function. The same regularization weight λ is used for both the quadratic and the robust objective functions.

3.2 Baseline Results

The regularization parameter λ is selected among a set of candidate values to achieve the best average end-point error (EPE) on the Middlebury training set. For the Charbonnier penalty function, the candidate set is [1, 3, 5, 8, 10] and 5 is optimal. The Charbonnier penalty uses $\epsilon = 0.001$ for both the data and the spatial term in Eq. 1. The Lorentzian uses $\sigma = 1.5$ for the data term, $\sigma = 0.03$ for the spatial term, and $\lambda = 0.06$. These parameters are fixed throughout the experiments, except where mentioned.

Table 1 summarizes the EPE results of the basic model with three different penalty functions on the Middlebury test set, along with the two top performers at the time of performing the evaluation (considering only published papers when the evaluation table was generated). Table 2 provides detailed results for each sequence. The classic formulations with two non-quadratic penalty functions (**Classic-C**) and (**Classic-L**) achieve competitive results despite their simplicity. The baseline optimization of **HS** and **BA (Classic-L)** results in significantly better accuracy than previously reported for these models ([Sun et al. 2008](#)). Note that the analysis also holds for the training set (Table 3).

Table 1 Models: average rank and end-point error (EPE) on the Middlebury test set using different penalty functions

	Avg. Rank	Avg. EPE
Classic-C	34.8	0.408
HS	49.0	0.501
Classic-L	42.7	0.530
Classic-C-brightness	N/A	0.726
HS-brightness	N/A	0.759
Classic-L-brightness	N/A	0.603
HS (Sun et al. 2008)	66.2	0.872
BA (Classic-L) (Sun et al. 2008)	59.6	0.746
Adaptive (Wedel et al. 2009)	28.5	0.401
Complementary OF (Zimmer et al. 2009)	31.6	0.485

Two state-of-the-art methods in Dec. 2010 are included for comparison. The ranking information was obtained at the writing of the paper (Sep. 2012). Please refer to Table 2 for the EPE results on each sequence

Because **Classic-C** performs quite well despite its simplicity, we set it as the baseline below. Note that our baseline implementation of **HS** has a lower average EPE than many more sophisticated methods. The **HS** implementation here incorporates many algorithmic and implementation details not present in the original **HS** method; the core idea of quadratic data and spatial terms however remains the same. In our naming convention, one can think of the **HS** method here as **Classic-Q**, meaning that it is the same as the **Classic-C** method except that the data and spatial penalty terms are quadratic.

4 Practices Explored

We now systematically vary the baseline approach by incorporating different ideas that have appeared in the literature, with the goal of illuminating which of these ideas are significant. This analysis is performed on the Middlebury training set by changing only *one property at a time*. Statistical significance is determined using a Wilcoxon signed rank test ([Wilcoxon 1945](#)) between each modified method and the baseline **Classic-C** method; a p value less than 0.05 indicates a significant difference. Each section below presents detailed comparisons of all these methods and then summarizes the results in a simple “take away message” about what we think are the “best practices” based on the data.

4.1 Image Pre-Processing

While it is common to talk about the brightness constancy assumption as a core feature of most optical flow algorithms, in practice many other constancy assumptions have been used. It is common, for example, to pre-filter the images in a variety of ways ranging from simple smoothing to edge

Table 2 Models: average end-point error (EPE) on the Middlebury optical flow benchmark (*test set*)

	Rank	Average	Army	Mequon	Schefflera	Wooden	Grove	Urban	Yosemite	Teddy
HS	49.0	0.501	0.12	0.25	0.45	0.24	0.95	0.83	0.24	0.93
Classic-C	34.8	0.408	0.10	0.23	0.45	0.20	0.88	0.47	0.16	0.77
Classic-L	42.7	0.530	0.10	0.24	0.47	0.21	0.92	1.23	0.20	0.87
HS-brightness	N/A	0.759	0.21	0.89	1.13	0.42	0.93	0.70	0.18	1.61
Classic-C-brightness	N/A	0.726	0.39	0.95	1.12	0.42	0.87	0.48	0.13	1.45
Classic-L-brightness	N/A	0.603	0.17	0.64	0.84	0.32	0.90	0.48	0.13	1.34
HS (Sun et al. 2008)	66.2	0.872	0.22	0.61	1.01	0.78	1.26	1.43	0.16	1.51
BA (Classic-L) (Sun et al. 2008)	59.6	0.746	0.18	0.58	0.95	0.49	1.08	1.43	0.15	1.11
Adaptive (Wedel et al. 2009)	28.5	0.401	0.09	0.23	0.54	0.18	0.88	0.50	0.14	0.65
Complementary OF (Zimmer et al. 2009)	31.6	0.485	0.10	0.20	0.35	0.19	0.87	1.46	0.11	0.60
NL-TV-NCC (Werlberger et al. 2010)	23.5	0.388	0.10	0.22	0.35	0.15	0.79	0.78	0.16	0.55
Classic++	32.7	0.406	0.09	0.23	0.43	0.20	0.87	0.47	0.17	0.79
Classic++Gradient	33.5	0.430	0.08	0.17	0.49	0.21	0.94	0.55	0.17	0.83
Classic+NL	17.2	0.319	0.08	0.22	0.29	0.15	0.64	0.52	0.16	0.49
Classic+NL-Full	17.5	0.316	0.08	0.24	0.28	0.15	0.63	0.49	0.16	0.50

The ranking information was determined at the writing of the paper (Sep. 2012)

Table 3 Pre-processing: average end-point error (EPE) on the Middlebury *training set* for the baseline method (**Classic-C**) using different image pre-processing techniques

	Avg. EPE	Significance	<i>p</i> value
Classic-C	0.298	–	–
HS	0.384	1	0.0078
Classic-L	0.319	1	0.0078
Classic-C-brightness	0.288	0	0.9453
HS-brightness	0.387	1	0.0078
Classic-L-brightness	0.325	0	0.2969
Gradient	0.305	0	0.4609
Gaussian + Dx + Dy	0.290	0	0.6406
Sobel edge magnitude (Vaudrey and Klette 2009)	0.417	1	0.0156
Laplacian (Lempitsky et al. 2010)	0.430	1	0.0078
Laplacian 1:1	0.301	0	0.6641
Gaussian pre-filtering ($\sigma = 0.5$)	0.281	0	0.5469
Texture 4:1	0.286	0	0.5312
Unnormalized texture	0.298	0	0.3750

Significance is always with respect to **Classic-C**. Please refer to Tables 4 and 5 for the detailed results on each training sequence
Bold entries highlight statistical significance

detection. For each method, we optimize the regularization parameter λ for the training sequences. The results are summarized in Table 3, with details of the methods applied to individual training sequences given in Tables 4 and 5. The baseline uses a non-linear pre-filtering of the images (ROF) to reduce the influence of illumination changes between frames

(Wedel et al. 2008b). Table 3 shows the effect of using no pre-processing, resulting in the standard brightness constancy model (***-brightness**). **Classic-C-brightness** actually achieves lower EPE on the training set than does **Classic-C** but significantly higher error on the test set (Table 1). This disparity suggests overfitting to the training data and leaves open the question as to whether the standard brightness constancy assumption, formulated robustly, may still compete with various types of filter/structure constancy given appropriate training data.

Simpler alternatives, such as filter response (or high-order) constancy (Brox et al. 2004; Bruhn & Weickert 2005; Sun et al. 2008) can serve the same purpose as ROF texture decomposition. A variety of pre-filters have been used in the literature, including derivative filters, Laplacians (Burt et al. 1982; Lempitsky et al. 2010), and Gaussians. Edges have also been emphasized using the Sobel edge magnitude (Vaudrey and Klette 2009).

Gradient only imposes constancy of the gradient vector at each pixel as proposed by Brox et al. (2004); i. e., it robustly penalizes the Euclidean distance between image gradients. We use central difference filters ($Dx = [-0.5 \ 0 \ 0.5]$ and $Dy = Dx^T$). **Gaussian+Dx+Dy** assumes separate brightness, horizontal derivative, and vertical derivative constancy. A weighted combination of robust functions applied to each term is used as by Sun et al. (2008). Neither of these methods differ significantly from the baseline texture decomposition (**Classic-C**). Two methods are significantly worse: the **Sobel edge magnitude** (Vaudrey and Klette 2009) and **Laplacian pre-filtering** (5×5) as used by Lempitsky et al. (2010). **Sobel**

Table 4 Models and pre-processing: average end-point error (EPE) on the Middlebury *training set* for the classical model and different penalty functions

	Average	Venus	Dimetrodon	Hydrangea	RubberWhale	Grove2	Grove3	Urban2	Urban3	Signif.	<i>p</i> value
Classic-C	0.298	0.281	0.152	0.165	0.093	0.158	0.627	0.348	0.562	–	–
Classic-C-brightness	0.288	0.268	0.166	0.215	0.134	0.146	0.584	0.352	0.437	0	0.9453
HS	0.384	0.337	0.219	0.189	0.118	0.204	0.688	0.463	0.853	1	0.0078
HS-brightness	0.387	0.335	0.226	0.252	0.154	0.185	0.639	0.564	0.743	1	0.0078
Classic-L	0.319	0.294	0.193	0.175	0.095	0.166	0.648	0.374	0.604	1	0.0078
Classic-L-brightness	0.325	0.292	0.207	0.274	0.145	0.158	0.588	0.451	0.484	0	0.2969

By default, the input sequences were preprocessed using ROF texture decomposition; “brightness” means no preprocessing is performed. The statistical significance is tested using the Wilcoxon signed rank test between each method and the baseline (**Classic-C**)

Bold entries highlight statistical significance

Table 5 Pre-processing: average end-point error (EPE) on the Middlebury *training set* for the baseline method (**Classic-C**) using different pre-processing techniques

	Average	Venus	Dimetrodon	Hydrangea	RubberWhale	Grove2	Grove3	Urban2	Urban3	Signif.	<i>p</i> value
Classic-C	0.298	0.281	0.152	0.165	0.093	0.158	0.627	0.348	0.562	–	–
Gradient	0.305	0.288	0.141	0.167	0.092	0.165	0.614	0.385	0.588	0	0.4609
Gaussian	0.281	0.268	0.146	0.226	0.141	0.137	0.582	0.335	0.413	0	0.5469
Gaussian + Dx + Dy	0.290	0.280	0.126	0.174	0.105	0.154	0.588	0.470	0.420	0	0.6406
Dx + Dy	0.301	0.286	0.122	0.166	0.099	0.161	0.616	0.443	0.518	0	1.0000
Sobel edge (Vaudrey and Klette 2009)	0.417	0.334	0.149	0.184	0.130	0.194	0.757	0.451	1.135	1	0.0156
Laplacian (Lempitsky et al. 2008)	0.430	0.374	0.170	0.176	0.096	0.175	0.756	0.464	1.232	1	0.0078
Laplacian 1:1	0.301	0.296	0.179	0.193	0.109	0.157	0.606	0.349	0.520	0	0.6641
Texture 4:1	0.286	0.271	0.159	0.175	0.100	0.154	0.587	0.349	0.490	0	0.5312
Unnormalized texture	0.298	0.279	0.152	0.166	0.092	0.158	0.623	0.348	0.563	0	0.3750

The regularization weight λ parameter was tuned for each method to achieve optimal performance. The statistical significance is tested using the Wilcoxon signed rank test between each method and the baseline (**Classic-C**)

Bold entries highlight statistical significance

edge magnitude appears to not work well on some of the sequences, particularly the synthetic ones, and may not be suitable for a general flow estimation method. **Laplacian** pre-filtering (5×5) as used by Lempitsky et al. (2010) produces good results on “RubberWhale”, but poor ones on the synthetic sequences. Note that the parameters for the FusionFlow method (Lempitsky et al. 2010) were mainly tuned using the “RubberWhale” sequence. The evaluation results suggest room for improving the FusionFlow method by a better pre-processing technique. **Gaussian pre-filtering** ($\sigma = 0.5$) performed well on the synthetic sequences, but poorly on real ones. Finally, the texture-structure blending ratio is 20:1 in Wedel et al. (2008b) but 4:1 in Werlberger et al. (2009). We find that (**Texture4:1**) performs better (but not significantly) on the synthetic sequences with a little degradation on the real ones. By default, the blended result from texture decomposition is normalized to $[-1, 1]$ by Wedel et al. (2008b) and $[0, 255]$ in our experiment. Not doing this normalization (**Unnormalized texture**) has little effect.

For the Laplacian pre-filtering, we find combining the filtered image with the original image, in the proportion 1:1, improves accuracy significantly (**Laplacian1:1**). Similar to the ROF texture decomposition, such an approach boosts the high frequency while suppressing the low frequency components that contain the lighting change.

Good Practices: Some form of image filtering is useful but simple derivative constancy is nearly as good as the more sophisticated texture decomposition method.

4.2 Coarse-to-Fine Estimation and Graduated Non-Convexity (GNC)

We vary the number of warping steps per pyramid level and find that **3 warping steps** gives similar results as using the baseline 10 (Table 6), except on “Urban3”, which is dominated by large motion and occlusions (see Table 7 for sequence-specific results). For the coarse-to-fine pyramid, Sun et al. (2008) use a downsampling factor of 0.8 during

Table 6 Model and methods: average end-point error (EPE) on the Middlebury *training set* for the baseline method (**Classic-C**) using different algorithm and modeling choices

	Avg. EPE	Significance	<i>p</i> value
Classic-C	0.298	—	—
3 warping steps	0.304	0	0.9688
Down-0.5	0.298	0	1.0000
w/o GNC	0.354	0	0.1094
Bilinear	0.302	0	0.1016
w/o TAVG	0.306	0	0.1562
Central derivative filter	0.300	0	0.7266
7-point derivative filter (Bruhn et al. 2005)	0.302	0	0.3125
Deriv-warp	0.297	0	0.9531
Bicubic-II	0.290	1	0.0391
Deriv-warp-II	0.287	1	0.0156
Warp-deriv-II	0.288	1	0.0391
C-L ($\lambda = 0.6$)	0.303	0	0.1562
L-C ($\lambda = 2$)	0.306	0	0.1562
GC-0.45 ($\lambda = 3$)	0.292	1	0.0156
GC-0.25 ($\lambda = 0.7$)	0.298	0	1.0000
MF 3×3	0.305	0	0.1016
MF 7×7	0.305	0	0.5625
2× MF	0.300	0	1.0000
5× MF	0.305	0	0.6875
w/o MF	0.352	1	0.0078
Classic++	0.285	1	0.0078

Please refer to Table 7 for the detailed results on each sequence
 Bold entries highlight statistical significance

non-convex optimization. A traditional downsampling factor of 0.5 (**Down-0.5**), however, has nearly identical performance. Note that a larger factor means that the pyramid levels are more similar in size and, for a pyramid with top bottom levels of the same size, results in more pyramid levels.

Previously, Brox et al. (2004) have reported that a downsampling factor of 0.95 produces much better results than 0.5. Note that for each iterative warping estimation step, Brox et al. use successive over-relaxation (SOR) to iteratively solve their linear system of equations and stop the iteration before convergence. With a downsampling factor of 0.95, they effectively increase the number of iterative warping steps performed by the algorithm, and this likely helps the overall algorithm converge. For our implementation, we solve the linear system of equations using the MATLAB built-in backslash function and obtain converged results for each iterative warping estimation step. Under such a setting, we find that the downsampling factor has little influence on the performance.

Removing the GNC procedure for the Charbonnier penalty function (**w/o GNC**) results in higher EPE on most sequences and higher energy on all sequences (Table 8). This suggests that the GNC method is helpful even for the convex Charbonnier penalty function due to the nonlinearity of the data term.

Good Practices: The downsampling factor does not matter when using a convex penalty; a standard factor of 0.5 is fine. Some form of GNC is useful even for a convex robust penalty like Charbonnier because of the nonlinear data term.

4.3 Interpolation Method and Derivatives

We find that the baseline bicubic interpolation is more accurate than bilinear (Table 6, **Bilinear**), as already reported in previous work (Wedel et al. 2008b). Removing temporal averaging of the gradients (**w/o TAVG**), using a **Central difference filter** $[-1 \ 0 \ 1]/2$, or using a **7-point derivative filter** $[-1 \ 9 \ -45 \ 0 \ 45 \ -9 \ 1]/60$ (Bruhn et al. 2005) all reduce accuracy compared to the baseline, but not significantly.

The baseline method computes the image derivative by first computing the derivative of the second image, warping the intermediate result toward the first image, and then averaging the warped result with the spatial derivative of the first image. Another approach is to first warp the second image toward the first image, compute the derivatives of the warped image, and then perform the temporal averaging with the spatial derivatives of the first image (Bruhn et al. 2005). We find the second approach produces similar results (**Deriv-warp**). However, the derivatives computed in either way are inconsistent with those implicitly interpolated by the bicubic interpolation. Bicubic interpolation interpolates not only the image but also the derivatives (Press et al. 2002). Because the MATLAB built-in function *interp2* is based on cubic convolution (Keys 1981) and does not provide the derivatives used in interpolation, we use the spline-based implementation by Press et al. (2002). With the new implementation (**Bicubic-II**), the three different ways to compute the derivatives give very similar EPE results, all better than the MATLAB built-in function. However, the one with consistent derivatives (**Bicubic-II**) gives the lowest energy solution, as shown in Table 9.

Good Practices: Use spline-based bicubic interpolation with a 5-point filter. Compute the derivatives during the interpolation to obtain the lowest energy solutions. Temporal averaging of the derivatives is probably worthwhile for a small computational expense.

4.4 Penalty Functions

We find that the convex Charbonnier penalty performs better than the more robust, non-convex Lorentzian on both the training and test sets. We test using the Charbonnier for the

Table 7 Model and methods: average end-point error (EPE) on the Middlebury *training set* for the baseline model (**Classic-C**) using different algorithm and modeling choices

	Average	Venus	Dimetrodon	Hydrangea	RubberWhale	Grove2	Grove3	Urban2	Urban3	Signif.	<i>p</i> value
Classic-C	0.298	0.281	0.152	0.165	0.093	0.158	0.627	0.348	0.562	–	–
3 warping steps	0.304	0.283	0.122	0.163	0.095	0.150	0.622	0.357	0.644	0	0.9688
Down-0.5	0.298	0.280	0.152	0.166	0.092	0.158	0.626	0.349	0.562	0	1.0000
Down-0.95	0.298	0.281	0.151	0.168	0.099	0.165	0.661	0.339	0.523	0	0.9375
w/o GNC	0.354	0.303	0.160	0.171	0.105	0.183	0.835	0.316	0.759	0	0.1094
Bilinear	0.302	0.284	0.144	0.167	0.099	0.160	0.637	0.363	0.563	0	0.1016
w/o TAVG	0.306	0.288	0.149	0.167	0.093	0.163	0.647	0.345	0.593	0	0.1562
Central	0.300	0.272	0.156	0.169	0.092	0.159	0.608	0.349	0.597	0	0.7266
7-point (Bruhn et al. 2005)	0.302	0.282	0.168	0.171	0.091	0.163	0.601	0.360	0.584	0	0.3125
Deriv-warp	0.297	0.283	0.153	0.165	0.092	0.159	0.636	0.333	0.552	0	0.9531
Bicubic-II	0.290	0.276	0.132	0.152	0.083	0.142	0.624	0.338	0.571	1	0.0391
Deriv-warp-II	0.287	0.264	0.155	0.152	0.085	0.145	0.616	0.333	0.546	1	0.0156
Warp-deriv-II	0.288	0.267	0.155	0.151	0.085	0.147	0.630	0.328	0.542	1	0.0391
C-L ($\lambda = 0.6$)	0.303	0.290	0.158	0.171	0.094	0.158	0.611	0.367	0.579	0	0.1562
L-C ($\lambda = 2$)	0.306	0.281	0.174	0.173	0.096	0.164	0.662	0.343	0.557	0	0.1562
GC-0.45 ($\lambda = 3$)	0.292	0.280	0.145	0.165	0.092	0.154	0.612	0.340	0.546	1	0.0156
GC-0.25 ($\lambda = 0.7$)	0.298	0.283	0.128	0.169	0.094	0.150	0.617	0.353	0.594	0	1.0000
MF 3×3	0.305	0.287	0.155	0.168	0.094	0.162	0.616	0.372	0.583	0	0.1016
MF 7×7	0.305	0.281	0.152	0.173	0.095	0.174	0.676	0.330	0.557	0	0.5625
$2 \times$ MF	0.300	0.279	0.152	0.167	0.093	0.163	0.650	0.339	0.555	0	1.0000
$5 \times$ MF	0.305	0.278	0.152	0.171	0.093	0.172	0.682	0.329	0.561	0	0.6875
w/o MF	0.352	0.307	0.168	0.199	0.113	0.217	0.705	0.423	0.684	1	0.0078
Classic++	0.285	0.271	0.128	0.153	0.081	0.139	0.614	0.336	0.555	1	0.0078

The statistical significance is tested using the Wilcoxon signed rank test between each method and the baseline (**Classic-C**)

Bold entries highlight statistical significance

Table 8 Energy ($\times 10^6$, Eq. 1) for the optical flow fields computed on the Middlebury *training set*, evaluated using convolution-based bicubic interpolation (Keys 1981)

	Sum	Venus	Dimetrodon	Hydrangea	RubberWhale	Grove2	Grove3	Urban2	Urban3
Classic-C	9.388	0.589	0.748	0.866	0.502	1.816	2.317	1.126	1.424
w/o GNC	9.689	0.593	0.750	0.870	0.506	1.845	2.518	1.142	1.465
w/o MF	8.044	0.517	0.701	0.668	0.449	1.418	1.830	1.066	1.395

Note that **Classic-C** uses graduated non-convexity (GNC), which reduces the energy, and median filtering, which increases it

Table 9 Energy ($\times 10^6$, Eq. 1) for the optical flow fields computed on the Middlebury *training set*, evaluated using spline-based bicubic interpolation (Press et al. 2002)

	Sum	Venus	Dimetrodon	Hydrangea	RubberWhale	Grove2	Grove3	Urban2	Urban3
Bicubic-II	8.761	0.552	0.734	0.835	0.481	1.656	2.167	1.061	1.275
Deriv-warp	8.917	0.559	0.745	0.840	0.484	1.682	2.201	1.073	1.333
Warp-deriv	9.035	0.563	0.745	0.845	0.486	1.694	2.238	1.117	1.347

Note the derivatives consistent with the interpolation method (**Bicubic-II**) produce the lowest energy solution

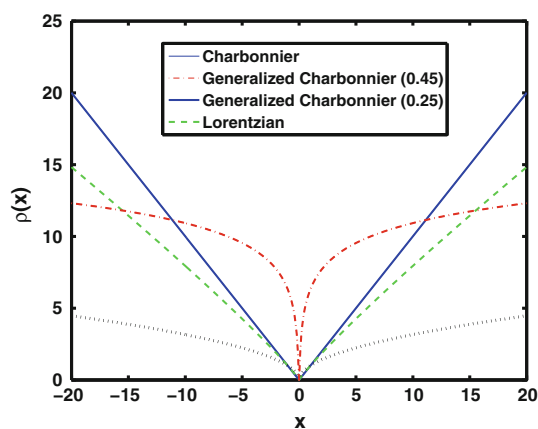


Fig. 2 Different penalty functions for the spatial terms: Charbonnier ($\epsilon = 0.001$), generalized Charbonnier ($a = 0.45$ and $a = 0.25$), and Lorentzian ($\sigma = 0.03$)

data term and Lorentzian for the spatial term (**C–L**) and vice versa (**L–C**). The two approaches perform better than using the Lorentzian for both terms but worse than using the Charbonnier for both terms.

One reason might be that non-convex functions are more difficult to optimize, causing the optimization scheme to find a poor local optimum. Another reason might be that the MAP estimator actually favors the “wrong” penalty functions (Nikolova 2007; Schmidt et al. 2010).

We investigate a generalized Charbonnier penalty function $\rho(x) = (x^2 + \epsilon^2)^a$ that is equal to the Charbonnier penalty when $a = 0.5$, and non-convex when $a < 0.5$ (see Fig. 2). We optimize the regularization parameter λ again. We find a slightly non-convex penalty with $a = 0.45$ (**GC-0.45**) performs consistently better than the Charbonnier penalty, whereas more non-convex penalties (**GC-0.25** with $a = 0.25$) show no improvement.

Good Practices: The less-robust Charbonnier is preferable to the highly non-convex Lorentzian and a slightly non-convex penalty function (**GC-0.45**) is better still.

4.5 Median Filtering

Figure 3 illustrates the median filtering step within the coarse-to-fine incremental estimation process. The baseline 5×5 median filter (**MF** 5×5) is better than both **MF** 3×3 (Wedel et al. 2008b) and **MF** 7×7 , but the difference is not significant (Table 6). When we perform 5×5 median filtering twice ($2 \times \text{MF}$) or five times ($5 \times \text{MF}$) per warping step, the results are worse. Finally, removing the median filtering step (**w/o MF**) makes the computed flow significantly less accurate with larger outliers as shown in Table 6 and Fig. 4.

One interesting result with **HS** is that repeatedly applying median filtering (20 times) at every warping step improves the **HS** formulation and the improvement is statistically significant (**HS** $20 \times \text{MF}$ in Table 10).

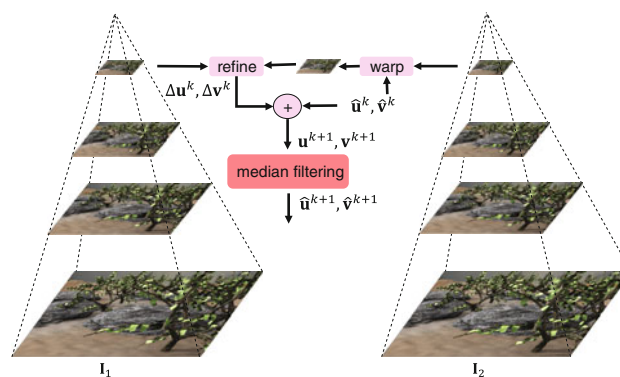


Fig. 3 The median filtering is performed after every incremental warping step (i.e., once at every image pyramid level). The output of the median filtering is upsampled and used as the initial estimate for the next larger pyramid level



Fig. 4 Estimated flow fields on sequence “RubberWhale” using **Classic-C** with and without (**w/o MF**) the median filtering step. **a** (**w/ MF**) energy 502, 387, **b** (**w/o MF**) energy 449, 290, **c** color key (Baker et al. 2007). The median filtering step helps reach a solution free from outliers but with a higher energy. The flow fields have been normalized by their maximum magnitude resulting in different contrasts. The outliers in the result without median filtering (**b**) make the flow appear lower contrast

Good Practices: Median filtering the intermediate flow results once after every warping iteration is the single most important implementation detail here; 5×5 is a good filter size.

4.6 Best Practices

Combining the analysis above into a single approach means modifying the baseline to use the slightly non-convex generalized Charbonnier and the spline-based bicubic interpolation. This leads to a statistically significant improvement over the baseline (Table 6, **Classic++**). This method is directly descended from **HS** and **BA**, yet updated with the current best optimization practices known to us. This simple method ranks 32th out of 73 methods in both EPE and AAE on the Middlebury test set at the writing of the paper (Sep. 2012). However, as we will see soon, this method is somehow not “simple”. Instead of the original objective, a different objective is being optimized with the median filtering step. The same is true for the reported results of both **HS** and **BA**.

Table 10 Additional results for **HS**: average end-point error (EPE) on the Middlebury *training set*

	Average	Venus	Dimetrodon	Hydrangea	RubberWhale	Grove2	Grove3	Urban2	Urban3	Signif.	<i>p</i> value
HS	0.384	0.337	0.219	0.189	0.118	0.204	0.688	0.463	0.853	–	–
HS 20× MF	0.365	0.299	0.214	0.184	0.104	0.196	0.699	0.431	0.792	1	0.0469

The statistical significance is tested using the Wilcoxon signed rank test between each method and **HS**
 Bold entries highlight statistical significance

5 Models Underlying Median Filtering

Our analysis reveals the practical importance of median filtering during optimization. This effectively *denoises* the intermediate flow fields, preventing gross outliers, and making even non-robust methods like **HS** more robust. We ask whether there is a *principle* underlying this heuristic?

One interesting observation is that flow fields obtained with median filtering have substantially *higher* energy than those without (Table 8; Fig. 4). If the median filter is helping to optimize the objective, it should lead to lower energies. Higher energies and more accurate estimates suggest that incorporating median filtering changes the objective function being optimized.

The insight that follows from this is that the median filtering heuristic is related to the minimization of an objective function that differs from the classical one. In particular the optimization of Eq. 1, with interleaved median filtering, approximately minimizes

$$E(\mathbf{u}, \mathbf{v}) = \sum_{i,j} \left\{ \rho_D(I_1(i, j) - I_2(i + u_{i,j}, j + v_{i,j})) + \lambda[\rho_S(u_{i,j} - u_{i+1,j}) + \rho_S(u_{i,j} - u_{i,j+1}) + \rho_S(v_{i,j} - v_{i+1,j}) + \rho_S(v_{i,j} - v_{i,j+1})] \right\} + \lambda_N \sum_{i,j} \sum_{(i',j') \in \mathcal{N}_{i,j}} (|u_{i,j} - u_{i',j'}| + |v_{i,j} - v_{i',j'}|), \quad (2)$$

where $\mathcal{N}_{i,j}$ is the set of neighbors of pixel (i, j) in a possibly large area and λ_N is a scalar weight. The term in braces is the same as the flow energy from Eq. 1, while the last term is new. This *non-local* term (Buades et al. 2005; Gilboa and Osher 2008) imposes a particular smoothness assumption within a specified region of the flow field.¹ Here we take this term to be a 5×5 rectangular region to match the size of the median filter in **Classic-C**. Figure 5 shows the neighborhood for the standard pairwise model and the non-local term.

It is usually difficult to directly optimize the objective (2) with a large spatial term. A common practice is to relax the objective with an auxiliary flow field as

¹ Bruhn et al. (2005) also integrated information over a local region in a global method but did so for the data term.

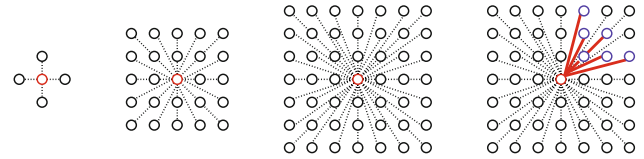


Fig. 5 From left to right, neighborhood structure for the center (red) pixel for the standard pairwise model, the unweighted non-local model, the unweighted non-local model with a larger neighborhood, and the weighted non-local model. The standard pairwise model connects a center pixel with its nearest neighbors, while the non-local term connects a pixel with many pixels in a large spatial neighborhood. By assigning larger weights (thicker red edges) to neighbors that are more likely to be on the same surface (blue circles), the weighted non-local model incorporates spatial scene structure information

$$E_A(\mathbf{u}, \mathbf{v}, \hat{\mathbf{u}}, \hat{\mathbf{v}}) = \sum_{i,j} \left\{ \rho_D(I_1(i, j) - I_2(i + u_{i,j}, j + v_{i,j})) + \lambda[\rho_S(u_{i,j} - u_{i+1,j}) + \rho_S(u_{i,j} - u_{i,j+1}) + \rho_S(v_{i,j} - v_{i+1,j}) + \rho_S(v_{i,j} - v_{i,j+1})] \right\} + \lambda_C(\|\mathbf{u} - \hat{\mathbf{u}}\|^2 + \|\mathbf{v} - \hat{\mathbf{v}}\|^2) + \lambda_N \sum_{i,j} \sum_{(i',j') \in \mathcal{N}_{i,j}} (|\hat{u}_{i,j} - \hat{u}_{i',j'}| + |\hat{v}_{i,j} - \hat{v}_{i',j'}|), \quad (3)$$

where $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ denote an auxiliary flow field and λ_C is a scalar weight. A third (coupling) term encourages $\hat{\mathbf{u}}$, $\hat{\mathbf{v}}$ and \mathbf{u} , \mathbf{v} to be the same (cf. Wedel et al. 2009; Zach et al. 2007). Here the notation implies a pixelwise sum of squared errors between the auxiliary and main flow fields.

The connection to median filtering (as a denoising method) derives from the fact that there is a direct relationship between the median and L1 minimization. Consider a simplified version of Eq. 3 with just the coupling and non-local terms, where

$$E(\hat{\mathbf{u}}) = \lambda_C \|\mathbf{u} - \hat{\mathbf{u}}\|^2 + \lambda_N \sum_{i,j} \sum_{(i',j') \in \mathcal{N}_{i,j}} |\hat{u}_{i,j} - \hat{u}_{i',j'}|. \quad (4)$$

While minimizing this is similar to median filtering \mathbf{u} , there are two differences. First, the non-local term minimizes the L1 distance between the central value and all flow values in its neighborhood except itself. Second, Eq. 4 incorporates information about the data term through the coupling equation; median filtering the flow ignores the data term.

The formal connection between Eq. 4 and median filtering² is provided by Li and Osher (2009) who show that minimizing Eq. 4 is related to a different median computation

$$\hat{u}_{i,j}^{(k+1)} = \text{median}(\text{Neighbors}^{(k)} \cup \text{Data}) \quad (5)$$

where $\text{Neighbors}^{(k)} = \{\hat{u}_{i',j'}^{(k)}\}$ for $(i', j') \in \mathcal{N}_{i,j}$ and $\hat{\mathbf{u}}^{(0)} = \mathbf{u}$ as well as

$$\text{Data} = \{u_{i,j}, u_{i,j} \pm \frac{\lambda_N}{\lambda_C}, u_{i,j} \pm \frac{2\lambda_N}{\lambda_C}, \dots, u_{i,j} \pm \frac{|\mathcal{N}_{i,j}|\lambda_N}{2\lambda_C}\},$$

where $|\mathcal{N}_{i,j}|$ denotes the (even) number of neighbors of (i, j) . Note that the set of “data” values is balanced with an equal number of elements on either side of the value $u_{i,j}$ and that information about the data term is included through $u_{i,j}$. Repeated application of Eq. 5 converges rapidly (Li and Osher 2009).

Observe that, as λ_N/λ_C increases, the weighted data values on either side of $u_{i,j}$ move away from the values of Neighbors and cancel each other out. As this happens, Eq. 5 approximates the median at the first iteration

$$\hat{u}_{i,j}^{(1)} \approx \text{median}(\text{Neighbors}^{(0)} \cup \{u_{i,j}\}). \quad (6)$$

Equation 3 thus combines the original objective with an approximation to the median, the influence of which is controlled by λ_N/λ_C . Note in practice the weight λ_C on the coupling term is usually small or is steadily increased from small values (Wedel et al. 2008b; Zach et al. 2007). We optimize the new objective (3) by alternately minimizing

$$\begin{aligned} E_O(\mathbf{u}, \mathbf{v}) = \sum_{i,j} \Big\{ & \rho_D(I_1(i, j) - I_2(i + u_{i,j}, j + v_{i,j})) \\ & + \lambda[\rho_S(u_{i,j} - u_{i+1,j}) + \rho_S(u_{i,j} - u_{i,j+1}) \\ & + \rho_S(v_{i,j} - v_{i+1,j}) + \rho_S(v_{i,j} - v_{i,j+1})] \Big\} \\ & + \lambda_C(\|\mathbf{u} - \hat{\mathbf{u}}\|^2 + \|\mathbf{v} - \hat{\mathbf{v}}\|^2) \end{aligned} \quad (7)$$

and

$$\begin{aligned} E_M(\hat{\mathbf{u}}, \hat{\mathbf{v}}) = & \lambda_C(\|\mathbf{u} - \hat{\mathbf{u}}\|^2 + \|\mathbf{v} - \hat{\mathbf{v}}\|^2) \\ & + \lambda_N \sum_{i,j} \sum_{(i',j') \in \mathcal{N}_{i,j}} (|\hat{u}_{i,j} - \hat{u}_{i',j'}| + |\hat{v}_{i,j} - \hat{v}_{i',j'}|). \end{aligned} \quad (8)$$

We find that optimization of the coupled set of equations is superior in terms of EPE performance than optimization of the objective in Eq. 2.

The alternating optimization strategy first holds $\hat{\mathbf{u}}, \hat{\mathbf{v}}$ fixed and minimizes Eq. 7 w.r.t. \mathbf{u}, \mathbf{v} . Then, with \mathbf{u}, \mathbf{v} fixed, we minimize Eq. 8 w.r.t. $\hat{\mathbf{u}}, \hat{\mathbf{v}}$. Note that Eqs. 4 and 8 can be minimized by repeated application of Eq. 5; we use this approach with five iterations. We perform 10 steps of alternating optimizations at every pyramid level and change λ_C logarithmically from 10^{-4} to 10^2 . During the first and second GNC

Table 11 Average end-point error (EPE) on the Middlebury *training set* is shown for the new model with alternating optimization (**Classic-C-A**)

	Avg. EPE	Significance	<i>p</i> value
Classic-C	0.298	–	–
Classic-C-A	0.305	0	0.8125
Classic-C-A-noRep	0.309	0	0.5781
Classic-C-A-II	0.296	0	0.7188
Classic-C-A-CGD	0.305	0	0.5625

Please refer to Table 13 for the detailed EPE results on each training sequence

stages, we set \mathbf{u}, \mathbf{v} to be $\hat{\mathbf{u}}, \hat{\mathbf{v}}$ after every warping step (this replacement step helps reach solutions with lower energy and EPE than without performing this step; see **Classic-C-A-noRep** in Tables 11, 12). In the end, we take $\hat{\mathbf{u}}, \hat{\mathbf{v}}$ as the final flow field estimate. The other parameters are $\lambda = 5$, $\lambda_N = 1$.

Alternately optimizing this new objective function (**Classic-C-A**) leads to similar results as the baseline **Classic-C** (Tables 11, 13). We also compare the energy of these solutions using the new objective and find the alternating optimization produces the lowest energy solutions, as shown in Table 12.

We find that approximately optimizing the new objective by changing λ_C logarithmically from 10^{-4} to 10^{-1} has slightly better EPE results but higher energy solutions (**Classic-C-A-II**). We also try replacing the absolute value by the Charbonnier penalty function and using the conjugate gradient descent method (<http://www.gaussianprocess.org/gpml/code/matlab/util/minimize.m>; last accessed 24 July 2013) to solve Eq. 4 but obtain results with slightly worse EPE performance and higher energy.

In summary, we show that the heuristic median filtering step in **Classic-C** can now be viewed as energy minimization of a new objective with a non-local term. The explicit formulation emphasizes the value of robustly integrating information over large neighborhoods and enables the improved model described below.

6 Improved Model

By formalizing the median filtering heuristic as an explicit objective function, we can find ways to improve it. While median filtering in a large neighborhood has advantages as we have seen, it also has problems. A neighborhood centered on a corner or thin structure is dominated by the surround and computing the median results in oversmoothing as illustrated in Fig. 1.

² Hsiao et al. (2003) established the connection in a slightly different way.

Table 12 Energy ($\times 10^6$, Eq. 3) for the computed flow fields on the Middlebury *training set*

	Sum	Venus	Dimetrodon	Hydrangea	RubberWhale	Grove2	Grove3	Urban2	Urban3
Classic-C	13.013	0.817	0.903	1.202	0.674	2.166	3.144	1.954	2.153
Classic-C w/o MF	14.629	0.886	0.945	1.299	0.725	2.315	3.513	2.234	2.712
Classic-C-A	12.489	0.784	0.889	1.139	0.666	2.064	2.976	1.922	2.049
Classic-C-A-noRep	13.076	0.790	0.894	1.165	0.670	2.092	3.143	2.005	2.317
Classic-C-A-II	13.308	0.830	0.915	1.235	0.686	2.223	3.247	1.990	2.182
Classic-C-A-CGD	13.466	0.833	0.909	1.224	0.674	2.213	3.357	2.020	2.236

The alternating optimization strategy (**Classic-C-A**) produces the lower energy solutions than the median filtering heuristic

Table 13 Average end-point error (EPE) on the Middlebury *training set* for the proposed new objective with the non-local term and alternating optimization (**Classic-C-A**) and its improved models

	Average	Venus	Dimetrodon	Hydrangea	RubberWhale	Grove2	Grove3	Urban2	Urban3	Signif.	p value
Classic-C	0.298	0.281	0.152	0.165	0.093	0.158	0.627	0.348	0.562	–	–
Classic-C-A	0.305	0.281	0.140	0.159	0.092	0.167	0.676	0.334	0.594	0	0.8125
Classic-C-A-noRep	0.309	0.279	0.139	0.161	0.093	0.157	0.653	0.370	0.619	0	0.5781
Classic-C-A-II	0.296	0.278	0.153	0.166	0.091	0.168	0.656	0.329	0.531	0	0.7188
Classic-C-A-CGD	0.305	0.281	0.148	0.161	0.093	0.159	0.697	0.344	0.560	0	0.5625

The statistical significance is tested using the Wilcoxon signed rank test between each method and the baseline (**Classic-C**)

Examining the non-local term suggests a solution. For a given pixel, if we know which other pixels in the area belong to the same surface, we can weight them more highly. The modification to the objective function is achieved by introducing a weight into the non-local term (Buades et al. 2005; Gilboa and Osher 2008):

$$\sum_{i,j} \sum_{(i',j') \in N_{i,j}} w_{i,j}^{i',j'} (|\hat{u}_{i,j} - \hat{u}_{i',j'}| + |\hat{v}_{i,j} - \hat{v}_{i',j'}|), \quad (9)$$

where $w_{i,j}^{i',j'}$ represents how likely pixel i', j' is to belong to the same surface as i, j .

Of course, we do not know $w_{i,j}^{i',j'}$, but can approximate it. We draw ideas from Sand and Teller (2008); Xiao et al. (2006); Yoon and Kweon (2006) to define the weights according to their spatial distance, their color-value distance, and their occlusion state as

$$w_{i,j}^{i',j'} \propto \exp \left\{ -\frac{|i-i'|^2 + |j-j'|^2}{2\sigma_1^2} - \frac{|\mathbf{I}(i,j) - \mathbf{I}(i',j')|^2}{2\sigma_2^2 n_c} \right\} \frac{o(i',j')}{o(i,j)}, \quad (10)$$

where $\mathbf{I}(i, j)$ is the color vector in the Lab space, n_c is the number of color channels, $\sigma_1 = 7$, $\sigma_2 = 7$, and the occlusion variable $o(i, j)$ is calculated using Eq. 22 in Sand and Teller (2008) as

$$o(i, j) = \exp \left\{ -\frac{d^2(i, j)}{2\sigma_d^2} - \frac{(I(i, j) - I(i + u_{i,j}, j + v_{i,j}))^2}{2\sigma_e^2} \right\}, \quad (11)$$

where $d(i, j)$ is the one-sided divergence function, defined as

$$d(i, j) = \begin{cases} \text{div}(i, j), & \text{div}(i, j) < 0 \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

in which the flow divergence $\text{div}(i, j)$ is

$$\text{div}(i, j) = \frac{\partial}{\partial x} u(i, j) + \frac{\partial}{\partial y} v(i, j), \quad (13)$$

where $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ are respectively the horizontal and vertical flow derivatives. The occlusion variable $o(i, j)$ is near zero for occluded pixels and near one for non-occluded pixels. We set the parameters in Eq. 11 as $\sigma_d = 0.3$ and $\sigma_e = 20$; this is the same as Sand and Teller (2008). Note that the occlusion state nonlinearly depends on the unknown flow field and we calculate the occlusion state using the latest flow estimate.

Examples of such weights are shown for several 15×15 neighborhoods in Fig. 6; bright values indicate higher weights. Note the neighborhood labeled **d**, corresponding to the rifle. Since pixels on the rifle are in the minority, an unweighted median oversmooths (**Classic++** in Fig. 1). The weighted term instead robustly estimates the motion using values on the rifle. A closely related piece of work is by Ren (2008), who uses the intervening contour to define affinities among neighboring pixels for the local Lucas and Kanade (1981) method. However it only uses this scheme to estimate motion for sparse points and then interpolates the dense flow field.

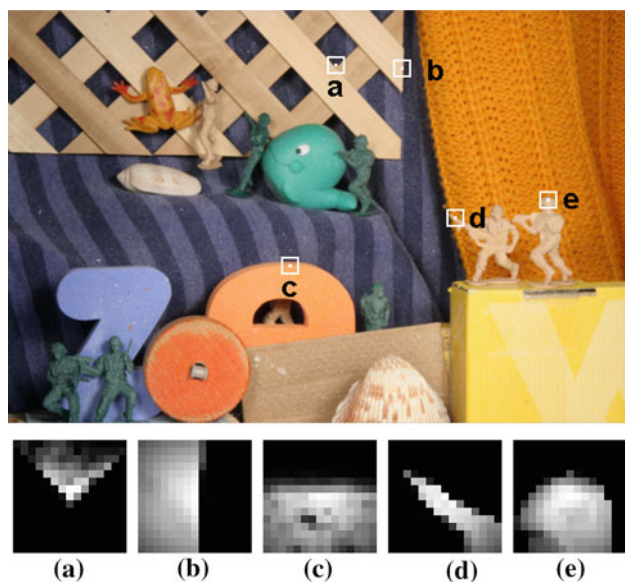


Fig. 6 Neighbor weights of the proposed weighted non-local term at different positions in the “Army” sequence. We use color, spatial distance, and occlusion cues to determine whether the neighboring pixels are likely to belong to the same surface. Among these cues, color is the most powerful (see Table 14 and text for an evaluation of the cues)

We approximately solve for $\hat{\mathbf{u}}$ (and similarly $\hat{\mathbf{v}}$) using the following weighted median problem

$$\min_{\hat{\mathbf{u}}_{i,j}} \sum_{(i',j') \in N_{i,j} \cup \{i,j\}} w_{i,j}^{i',j'} |\hat{\mathbf{u}}_{i,j} - \mathbf{u}_{i',j'}|, \quad (14)$$

using the formula (3.13) in Li and Osher (2009) for all the pixels (**Classic+NL-Full**). Note if all the weights are equal, the solution is just the median. In practice, we can adopt a fast version (**Classic+NL**) without performance loss: Given a current estimate of the flow, we detect motion boundaries using a Sobel edge detector and dilate these edges with a 5×5 mask to obtain flow boundary regions. In these regions we use the weighting in Eq. 10 in a 15×15 neighborhood. In the non-boundary regions, we use equal weights in a 5×5 neighborhood to compute the median.

To further reduce the computation, we can adopt a two-stage GNC process and perform three warping steps per pyramid level. This fast version (**Classic+NL-Fast**) has nearly the same overall performance, with a slight decline in performance on the “Urban3” sequence, which has large motions; with an iterative warping scheme, large motions require more iterations.

Tables 14 and 15 show that the weighted non-local term (**Classic+NL**) improves the accuracy on both the training and the test sets, especially in the motion boundary regions. Note that the fine detail of the “rifle” is preserved in Fig. 1e. At the writing of this paper (Sep. 2012), **Classic+NL** ranks 13th in both AAE and EPE. Figures 7 and 8 show some of the results on the Middlebury dataset.

Table 14 Average end-point error (EPE) on the Middlebury *training set* is shown for the improved model and its variants

	Avg. EPE	Significance	<i>p</i> value
Classic+NL	0.221	–	–
Classic+NL-Full	0.222	0	0.8203
Classic+NL-Fast	0.221	0	0.3125
RGB	0.240	1	0.0156
HSV	0.231	1	0.0312
LUV	0.226	0	0.5625
Gray	0.253	1	0.0078
w/o color	0.283	1	0.0156
w/o occ	0.226	0	0.1250
w/o spa	0.223	0	0.5625
$\sigma_2 = 5$	0.221	0	1.0000
$\sigma_2 = 10$	0.224	0	0.2500
$\lambda = 1$	0.236	0	0.1406
$\lambda = 9$	0.244	0	0.1016
11×11	0.223	0	0.5938
19×19	0.220	0	0.8750

Please refer to Table 16 for the detailed results

Bold entries highlight statistical significance

Table 15 Average end-point error (EPE) on the Middlebury *test set* for the **Classic++** model with two different preprocessing techniques and its improved model

	Avg. rank	Avg. EPE	Avg. EPE near boundary
Classic++	32.7	0.406	0.980
Classic++Gradient	33.5	0.430	1.042
Classic+NL	17.2	0.319	0.689
Classic+NL-Full	17.5	0.316	0.676

Please refer to Table 2 for the detailed EPE results

We study some variants of the weighted non-local term (**Classic+NL**). Tables 14 and 16 show the importance of each term in determining the weight and influence of the parameter setting on the final results. Using different color spaces results in some performance decline. Using grayscale pixel values (**Gray**) or not using the static image information (**w/o color**) results in significant degradation in performance. Without occlusion (**w/o occ**) or spatial distance (**w/o spa**) cues does not degrade the performance significantly. The method is robust to the setting of σ_2 for the color cue and 5 and 10 perform similarly as the default 7. The default λ is 3, while 1 and 9 result in some loss in performance. We also study the maximum size of the neighborhood for the non-local term and find 11×11 gives similar performance while 19×19 is slightly better.

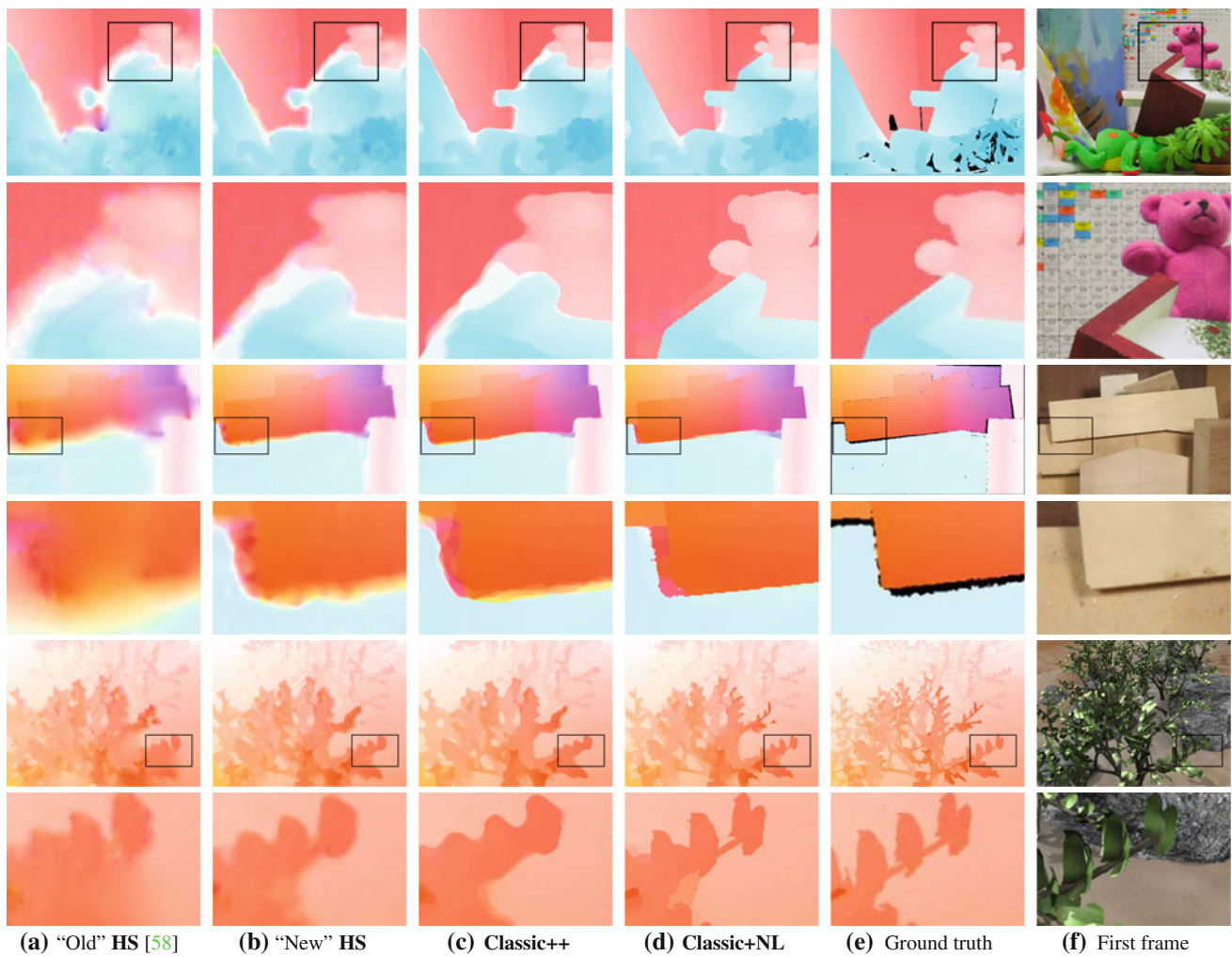


Fig. 7 Results on the Middlebury test set. *Top to bottom*: "Teddy", "Wooden", and "Grove". **Classic+NL** uses information from the color to detect and preserve fine motion details. Note that the ground

truth visualization from the Middlebury website has been compressed and has lower quality than the actual ground truth

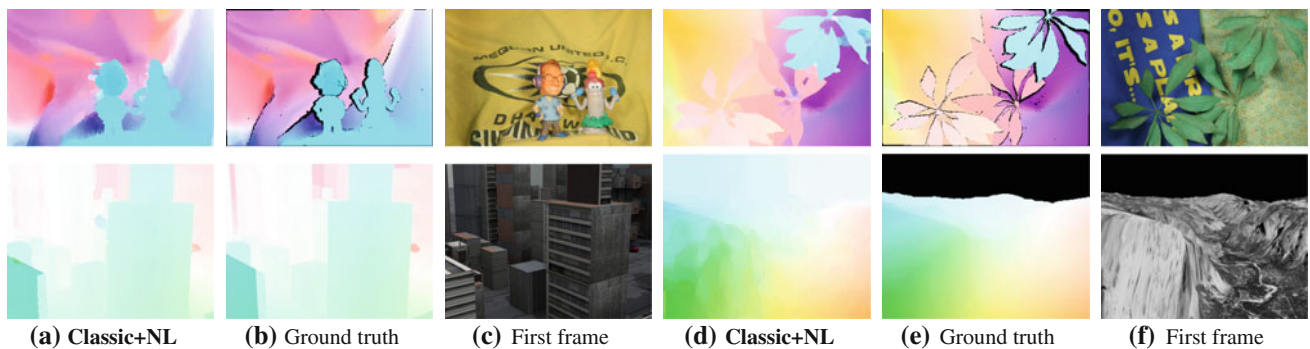


Fig. 8 Results on other Middlebury test sequences. *Top left* "Mequon"; *top right* "Schefflera"; *bottom left* "Urban"; *bottom right* "Yosemite"

6.1 Closely-Related Work

Werlberger et al. (2010) independently propose a non-local term for optical flow estimation and the spatial term is similar

to our non-local term. They use zero mean normalized cross correlation as the data term to deal with lighting changes. Their work is motivated by the success of the non-local regularization (Buades et al. 2005) in image restoration and

Table 16 Average end-point error (EPE) on the Middlebury *training set* for the proposed new objective with the weighted non-local term and its variants

	Average	Venus	Dimetrodon	Hydrangea	RubberWhale	Grove2	Grove3	Urban2	Urban3	Signif.	<i>p</i> value
Classic+NL	0.221	0.238	0.131	0.152	0.073	0.103	0.468	0.220	0.384	–	–
Classic+NL-Full	0.222	0.252	0.135	0.156	0.074	0.097	0.469	0.214	0.382	0	0.8203
Classic+NL-Fast	0.221	0.233	0.117	0.151	0.076	0.098	0.464	0.210	0.421	0	0.3125
RGB	0.240	0.243	0.131	0.155	0.081	0.109	0.501	0.236	0.468	1	0.0156
HSV	0.231	0.245	0.131	0.152	0.074	0.110	0.492	0.222	0.424	1	0.0312
LUV	0.226	0.241	0.131	0.149	0.074	0.104	0.460	0.223	0.427	0	0.5625
Gray	0.253	0.253	0.133	0.158	0.086	0.125	0.547	0.242	0.479	1	0.0078
w/o color	0.283	0.258	0.128	0.157	0.087	0.155	0.633	0.303	0.543	1	0.0156
w/o occ	0.226	0.243	0.131	0.152	0.073	0.103	0.488	0.230	0.386	0	0.1250
w/o spa	0.223	0.237	0.132	0.154	0.073	0.102	0.475	0.213	0.398	0	0.5625
$\sigma_2 = 5$	0.221	0.240	0.131	0.151	0.073	0.104	0.466	0.208	0.392	0	1.0000
$\sigma_2 = 10$	0.224	0.238	0.132	0.153	0.073	0.102	0.485	0.228	0.384	0	0.2500
$\lambda = 1$	0.236	0.245	0.151	0.164	0.080	0.120	0.430	0.243	0.459	0	0.1406
$\lambda = 9$	0.244	0.249	0.137	0.160	0.091	0.111	0.577	0.201	0.426	0	0.1016
11×11	0.223	0.240	0.131	0.151	0.074	0.103	0.451	0.234	0.397	0	0.5938
19×19	0.220	0.238	0.132	0.154	0.073	0.103	0.470	0.210	0.384	0	0.8750

The statistical significance is tested using the Wilcoxon signed rank test between each method and the baseline (**Classic+NL**)

Bold entries highlight statistical significance

stereo. Our work is inspired by the success of the heuristic median filtering step in flow estimation and we formalize the median filtering heuristic as a non-local regularization term. The use of the GPU and C++ makes their implementation faster than our implementation in MATLAB. **Classic+NL** has lower average EPE on the Middlebury test sequences; 0.319 versus 0.388 (cf. Table 2). Readers can visually compare the results of both methods on the Middlebury website.

6.2 Results on the MIT Dataset

To test the robustness of these models on other data, we applied **HS**, **Classic-C**, and **Classic+NL** to sequences from the MIT dataset (Liu et al. 2008), and compared the estimated flow fields to the human labeled ground truth. Note only five of the eight test sequences of Liu et al. (2008) are available on-line; these are tested here.

Figure 9 and Table 17 show the results on these sequences, which are very different in nature from the Middlebury set and include an outdoor scene as well as a scene of a fish tank. The results are compared with the CLG method (Bruhn et al. 2005) used by Liu et al. (2008). It is important to point out that the CLG method was tuned to obtain the optimal results on the test sequences. Our method had no such tuning and we used the same parameters as those used in all the other experiments. This suggests that training on the Middlebury data results in a method that generalizes to other sequences. The only place where this fails is on the “fish” sequence where

there is transparent motion in a liquid medium; the statistics in this sequence are very different from the Middlebury training data.

6.3 Performance on MPI Sintel and KITTI Datasets

We evaluate the methods above (corresponding to our publicly released code) on the MPI Sintel (Butler et al. 2012) and the KITTI (Geiger et al. 2012) datasets using the default parameter settings in our conference paper (Sun et al. 2010a). As summarized by Tables 17 and 19, the conclusions contradict our findings reported above. On the MPI Sintel dataset, **HS** outperforms **Classic++**, which in turn outperforms **Classic+NL-fast**. The only consistent result is **Classic+NL**, which achieves the best performance. On the KITTI dataset, **HS** outperforms **Classic+NL**.

We ask how these datasets differ from both Middlebury and the MIT dataset. What could lead to these inconsistent conclusions? One answer surprisingly lies in the unequal width and height of the images.

6.4 Asymmetric Pyramids for Wide-Aspect-Ratio Video

Our original implementation downsamples the image equally in the horizontal and vertical dimensions. The method automatically determines the number of pyramid levels using the smaller of the height and width of the input image. This

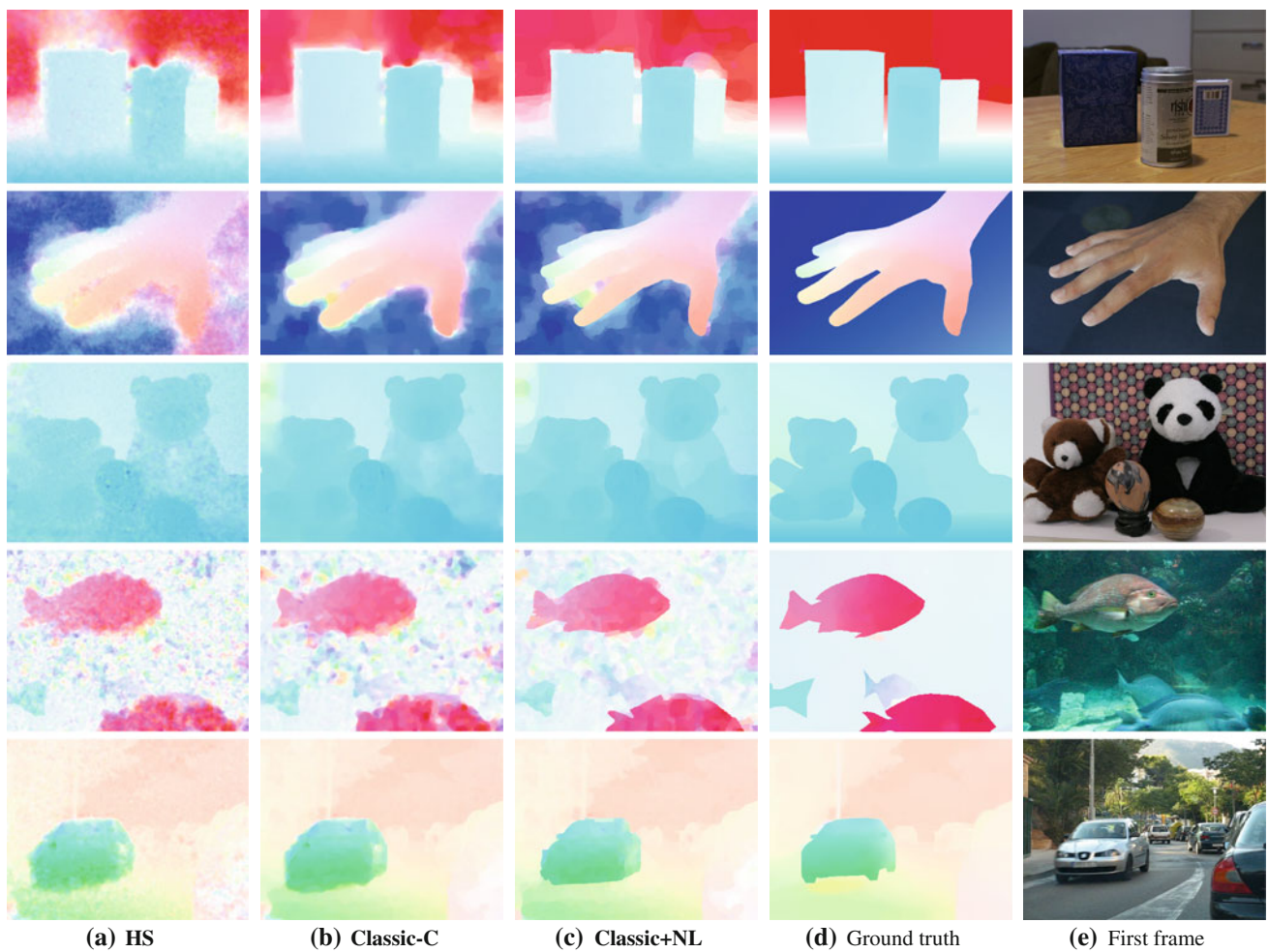


Fig. 9 Results on MIT sequences. *Top to bottom* “Table”, “Hand”, “Toy”, “Fish”, and “CameraMotion”

Table 17 Results on the MIT dataset (Liu et al. 2008)

	Average	Table	Hand	Toy	Fish	Camera Motion
CLG	1.239	0.976	4.181	0.456	0.196	0.385
HS	2.129	1.740	6.108	0.620	1.309	0.869
Classic-C	1.345	1.064	3.428	0.482	1.061	0.690
Classic+NL	1.106	0.91	2.75	0.487	0.772	0.611

Average end-point error (EPE). The CLG (Bruhn et al. 2005) method was tuned for each sequence (Liu et al. 2008)

scheme works well when the width-to-height ratio is close to 1, i.e., the Middlebury sequences. In contrast, the MPI Sintel images are $1,024 \times 436$ and the KITTI images are around $1,226 \times 370$. The small vertical dimension limits the height of the pyramid, but we find that the large horizontal dimension means that the sequences contain very large horizontal motions. As a result, at the top level of the pyramid, the horizontal motions can be much larger than a pixel.

To address this we can use an unequal downsampling factor in each direction to ensure that the motion at the top pyra-

Table 18 Average end-point error (EPE) on the MPI Sintel *training* set

	Classic+NL	Classic+NL-fast	Classic++	HS
Final	7.998	8.806	8.601	8.386
Clean	6.035	7.445	6.765	7.238
	Classic+NLP	Classic+NL-FastP	Classic++P	HSP
Final	7.459	7.379	7.928	8.291
Clean	4.942	5.078	5.363	6.899

mid level is small in both directions (or at least similar). For the MPI Sintel and KITTI data sets, we use a downsampling factor of 0.5 in the horizontal direction and determine the downsampling factor in the vertical direction and the pyramid level number, so that the size of top pyramid level is around 16×16 .

For MPI Sintel and KITTI this scheme results in a 7-level pyramid (instead of a 5-level pyramid in the standard symmetric scheme). This results in a significant improvement on both the the MPI Sintel and the KITTI data set, as summarized by Tables 18, 19 and 20. We denote the method with

Table 19 Average end-point error (EPE) on the MPI Sintel *test* set

	Classic+NL	Classic+NL-fast	Classic++	HS
Final	9.153	10.088	9.959	9.610
Clean	7.961	9.129	8.721	8.739
	Classic+NLP	Classic+NL-FastP	Classic++P	HSP
Final	8.291	8.439	8.836	9.377
Clean	6.731	6.940	7.209	8.309
	MDP-Flow2	FC-2Layers-FF	AHOF	
Final	8.445	8.137	11.927	
Clean	5.837	6.781	12.642	

Only published results are shown, including **MDP-Flow2** (Xu et al. 2012), **FC-2Layers-FF** (Sun et al. 2013) and **AHOF** (Werlberger et al. 2009)

Table 20 Percentage of pixels with EPE larger than 3 pixels in non-occluded (**Out-Noc**) and all **Out-All** areas and average EPE in non-occluded (**Avg-Noc**) and all **Avg-All** areas on the KITTI *test* set

Method	Out-Noc (%)	Out-All (%)	Avg-Noc (pixel)	Avg-All (pixel)
HS	19.92	28.86	5.8	11.7
Classic+NL	24.64	33.35	9.0	16.4
HSP	14.77	24.08	4.0	9.0
Classic+NL-FastP	12.42	22.27	3.2	7.8
Classic+NLP	10.60	20.66	2.8	7.2
Classic++P	10.16	20.29	2.6	7.1

the new asymmetric pyramid by adding an “P” at the end of the name.

On MPI Sintel, the results of the four methods are consistent with those on the Middlebury data set. Note that even **Classic++P** outperforms the previous **Classic+NL**. **Classic+NLP** outperforms **MDP-flow2** (Xu et al. 2012) on the final set, but not on the clean set. **MDP-flow2** uses feature matching to deal with fast moving objects. Feature matching tends to work well on the clean set, but not the final set due to motion and optical blur in the latter. Figure 10 shows an example visual comparison between results using **Classic+NLP** and **Classic+NL**. The asymmetric pyramid leads to significant improvement in areas that undergo large motions.

On the KITTI set, **Classic++P** performs best among all our tested methods, both in the training and the test sets. Note that the KITTI sequences have been collected on a moving vehicle in an urban environment. The flow fields tend to be smooth with few flow boundaries. The image-independent smoothness assumption in **Classic++P** is better suited to such data. Figure 11 shows some results for **Classic+NL-FastP** and **Classic+NL-fast**; note the dra-

matic improvement resulting from the asymmetric pyramid.

It is important to note that, apart from the change of pyramid method, all other parameters remain the same and are trained using the Middlebury training sequences.

6.5 Computational Time

Table 21 summarizes the running time of the evaluated methods on typical sequences from three different data sets in MATLAB on a 64-bit Linux desktop with 8GB memory. The additional cost from **HS** to **Classic++** comes from the GNC stage and the non-convex penalty function. The additional cost from **Classic++** to **Classic+NL** comes from the weighted median filtering step for detected motion boundaries. Applying the weighted median operation on all the pixels (**Classic+NL-Full**) increases the running time by more than three times with little performance gain. Using fewer iterations (**Classic+NL-Fast**) can significantly reduce the computational cost with little performance loss, especially on sequences with small motion. Note that we solve the weighted median problem at each pixel individually and do not reuse the sorting results from neighboring pixels. Future work should consider reformulating the weighted median filtering so that a convolution-type operation can be used to reduce the computational cost.

6.6 Limitations

Classic+NL produces larger errors in occlusion regions on some sequences, such as “Schefflera” shown in Fig. 12. The classical flow formulation assumes that every pixel at the current frame has a corresponding pixel at the next frame. However, this assumption breaks down in regions of occlusion. Pixels that are occluded by some foreground objects in one frame do not have corresponding pixels in the next, resulting in large errors with classical formulations. In contrast, a layered model (Wang and Adelson 1994) may provide a principled way to reason about occlusions. The motion model developed in this paper has enabled a recent layered approach (Sun et al. 2010b) to achieve a consistent improvement over the **Classic+NL** method, in particular near occlusion and motion boundary regions.

Small, fast moving objects also cause problems for the classical coarse-to-fine estimation used by **Classic+NL**, as shown in Fig. 10. The work by Brox and Malik (2011) on large displacement optical flow has inspired recent work (Chen et al. 2013; Steinbrücker et al. 2009; Xu et al. 2012) to embed feature matching into the coarse-to-fine estimation framework. Chen et al. (2013) show that, with proper initialization, **Classic+NL** can also handle large displacement optical flow on the Middlebury dataset.



Fig. 10 Example results on MPI Sintel dataset. From *top to bottom*: first frame, second frame, results by **Classic+NL** (5-level), results by **Classic+NLP** (7-level), and ground truth. The asymmetric pyramid leads to a significant improvements in large regions undergoing large motion (head of the dragon on the *left* and background on the *right*).

EPE results: “temple2” (*left*), 18.04 by **Classic+NL** (5-level) and 12.92 by **Classic+NLP** (7-level); “cave2” (*right*), 52.208 by **Classic+NL** (5-level) and 26.565 by **Classic+NLP** (7-level). Note that the estimated motion for fast-moving objects still contains large errors



Fig. 11 Example results on the KITTI dataset. From *top to bottom*: first frame, second frame, results by **Classic+NL-fast** (5-level), results by **Classic+NL-FastP** (7-level), and ground truth for the non-occluded feature points. EPE results in non-occluded sparse feature points: “000002”

(*left*), 12.124 by **Classic+NL-fast** (5-level) and 2.444 by **Classic+NL-FastP** (7-level); “000030” (*right*), 20.554 by **Classic+NL-fast** (5-level) and 0.615 by **Classic+NL-FastP** (7-level)

Table 21 Running time (in minutes) for computing one optical flow field from an image pair from different benchmark datasets using different methods in MATLAB on a 64-bit Linux desktop with 8GB memory

	Middlebury	MPI Sintel	KITTI
HS	1.62	1.8	2.56
Classic++	5.83	7.2	8.48
Classic+NL	9.81	14	14.78
C+NL-fast	1.8	2.5	2.89
C+NL-full	26.7	29	42

Used sequences: 640×480 Urban from Middlebury, 1024×436 alley_1 from MPI Sintel, and 1226×370 training image 0 from KITTI

7 Conclusions

When implemented using modern practices, classical optical flow formulations can produce fairly competitive results

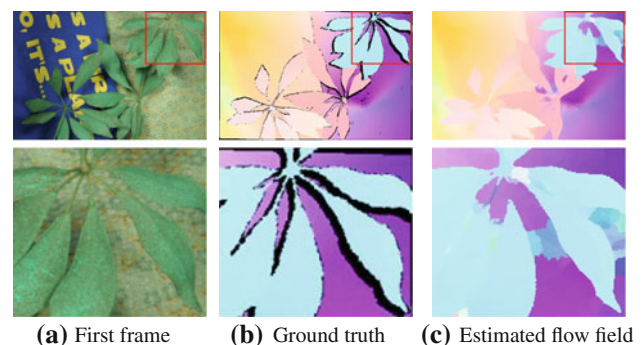


Fig. 12 Occlusions are not explicitly modeled by **Classic+NL** and may cause problems in the estimated flow field. Dark pixels in the ground truth indicate occlusions

on existing datasets. To understand the techniques that help such basic formulations work well, we quantitatively studied various aspects of flow approaches from the lit-

erature, including their implementation details. Among the best practices, we found that using median filtering to denoise the flow after every warping step is key to improving accuracy, but that this increases the energy of the final result. Exploiting connections between median filtering and L1-based denoising, we showed that algorithms relying on a median filtering step are approximately optimizing a different objective that regularizes the flow field over a large spatial neighborhood. Understanding this enables us to design and optimize improved models that weight the neighbors adaptively in an extended image region. The MATLAB code is publicly available at <http://www.cs.brown.edu/people/dqsun>; last accessed 24 July 2013.

There has been much debate about whether methods that perform well on Middlebury will generalize to other sequences. Here we tuned the parameters of the method on the Middlebury training set and tested on Middlebury, MIT HAMA, MPI Sintel, and KITTI. The conclusions on the Middlebury dataset are consistent with those on the MIT HAMA dataset. The one significant difference we found between Middlebury and the MPI Sintel and KITTI datasets was the aspect ratio of the images. This allowed us to make a change to the method by introducing a novel asymmetric image pyramid that downsamples more rapidly in the horizontal direction than the vertical direction. With only this change we found that our conclusions on Middlebury hold for MPI Sintel as well. The KITTI data set is somewhat different in nature and seems to favor methods with more spatial smoothing. As a result, the image-independent **Classic++**, which produces more smooth flow fields, performs slightly better than the image-dependent **Classic+NL**, with its sharp boundaries. It is open whether these conclusions will hold for data taken under totally different conditions, such as medical images. While the results on Middlebury generalize surprisingly well, we suspect that training the parameters for a specific dataset will improve results further.

Acknowledgments DS and MJB were supported by a gift from Intel Corp. and NSF CRCNS award IIS-0904875. DS was also supported by Kitware, Nvidia, Google, and the Intel Science and Technology Center for Visual Computing. We thank the CVPR and IJCV reviewers for constructive comments, especially the connection between our original “area” term and non-local regularization and test on additional datasets, P. Yadollahpour for his early work on implementing the HS method, S. Zuffi for suggesting the color version of the non-local term, J. Wulff for running some experiments, T. Brox, A. Wedel, and M. Werlberger for clarifying details about their papers, and D. Scharstein for maintaining the online optical flow benchmark.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

- Adato, Y., Zickler, T., & Ben-Shahar, O. (2011). A polar representation of motion and implications for optical flow. In *IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 1145–1152).
- Adelson, E. H., Anderson, C. H., Bergen, J. R., Burt, P. J., & Ogden, J. M. (1984). Pyramid methods in image processing. *RCA Engineer*, 29(6), 33–41.
- Álvarez, L., Castaño-Moraga, C. A., García, M., Krissian, K., Mazorra, L., Salgado, A., & Sánchez, J. (2007). *Symmetric optical flow*. EUROCAST (pp. 676–683). Springer, Berlin.
- Bab-Hadiashar, A., & Suter, D. (1998). Robust optic flow computation. *International Journal of Computer Vision*, 29(1), 59–77.
- Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M. J., & Szeliski, R. (2007). A database and evaluation methodology for optical flow. In *IEEE International Conference on Computer Vision*.
- Barron, J., Fleet, D., & Beauchemin, S. (1994). Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1), 43–77.
- Bergen, J., Anandan, P., Hanna, K., & Hingorani, R. (1992). Hierarchical model-based motion estimation. In *European Conference on Computer Vision* (Vol. 588, pp. 237–252).
- Black, M., & Jepson, A. (1996). Estimating optical-flow in segmented images using variable-order parametric models with local deformations. *IEEE Transaction on Pattern Analysis Machine Intelligence*, 18(10), 972–986.
- Black, M. J., & Anandan, P. (1996). The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63, 75–104.
- Blake, A., & Zisserman, A. (1987). *Visual reconstruction*. Cambridge, MA: The MIT Press.
- Brox, T., Bruhn, A., Papenberg, N., & Weickert, J. (2004). High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision* (Vol. IV, pp. 25–36).
- Brox, T., & Malik, J. (2011). Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Transaction on Pattern Analysis Machine Intelligence*, 33(3), 500–513.
- Bruhn, A., Weickert, J. (2005). Towards ultimate motion estimation: Combining highest accuracy with real-time performance. In *IEEE International Conference on Computer Vision* (Vol. 1, pp. 749–755).
- Bruhn, A., Weickert, J., & Schnörr, C. (2005). Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 61(3), 211–231.
- Buades, A., Coll, B., & Morel, J. (2005). A non-local algorithm for image denoising. In *IEEE International Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. 60–65).
- Burt, P. J., Yen, C., & Xu, X. (1982). Local correlation measures for motion analysis: A comparative study. In *Proceedings of IEEE Pattern Recognition and Image Processing* (pp. 269–274).
- Butler, D. J., Wulff, J., Stanley, G. B., & Black, M. J. (2012). A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision* (Vol. IV, pp. 611–625).
- Chen, Z., Jin, H., Lin, Z., Cohen, S., & Wu, Y. (2013). Large displacement optical flow from nearest neighbor fields. In *IEEE International Conference on Computer Vision and Pattern Recognition*.
- Chen, Z., Wu, Y., & Wang, J. (2012). Decomposing and regularizing sparse/non-sparse components for motion field estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 1176–1183).
- Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 3354–3361).

- Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis Machine Intelligence*, 6(6), 721–741.
- Gilboa, G., & Osher, S. (2008). Nonlocal operators with applications to image processing. *SIAM Multiscale Modeling and Simulation*, 7, 1005–1028.
- Glaer, F., Reynolds, G., & Anandan, P. (1983). Scene matching by hierarchical correlation. In *IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 432–441).
- Horn, B. (1986). *Robot vision*. Cambridge, MA: MIT Press.
- Horn, B., & Schunck, B. (1981). Determining optical flow. *Artificial Intelligence*, 16(1–3), 185–203.
- Hsiao, I., Rangarajan, A., & Gindi, G. (2003). A new convex edge-preserving median prior with applications to tomography. *IEEE Transactions on Medical Imaging*, 22(5), 580–585.
- Humayun, A., Mac Aodha, O., & Brostow, G. J. (2011). Learning to find occlusion regions. In *IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 2161–2168).
- Jia, K., Wang, X., & Tang, X. (2011). Optical flow estimation using learned sparse model. In *IEEE International Conference on Computer Vision* (pp. 2391–2398).
- Keys, R. G. (1981). Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(6), 1153–1160.
- Krähenbühl, P., & Koltun, V. (2012). Efficient nonlocal regularization for optical flow. In *European Conference on Computer Vision* (Vol. I, pp. 356–369).
- Lei, C., Yang, Y.-H. (2009). Optical flow estimation on coarse-to-fine region-trees using discrete optimization. In *IEEE International Conference on Computer Vision* (pp. 1562–1569).
- Lempitsky, V., Roth, S., Rother, C. (2008). FusionFlow: Discrete-continuous optimization for optical flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*.
- Lempitsky, V., Rother, C., Roth, S., & Blake, A. (2010). Fusion moves for Markov random field optimization. *IEEE Transaction on Pattern Analysis Machine Intelligence*, 32(8), 1392–1405.
- Li, Y., & Osher, S. (2009). A new median formula with applications to PDE based denoising. *Communications in Mathematical Sciences*, 7(3), 741–753.
- Lin, D., & Fisher, J. (2012). Low level vision via switchable Markov random fields. In *IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 2432–2439).
- Liu, C., Freeman, W. T., Adelson, E. H., & Weiss, Y. (2008). Human-assisted motion annotation. In *IEEE International Conference on Computer Vision and Pattern Recognition*.
- Lucas, B., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *International Joint Conferences on Artificial Intelligence* (pp. 674–679).
- Mac Aodha, O., Brostow, G. J., & Pollefeys, M. (2010). Segmenting video into classes of algorithm-suitability. In *IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 1778–1785).
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: W.H. Freeman.
- Nagel, H.-H., & Enkelmann, W. (1986). An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *IEEE Transaction on Pattern Analysis Machine Intelligence*, 8(5), 565–593.
- Nikolova, M. (2007). Model distortions in Bayesian MAP reconstruction. *AIMS Journal on Inverse Problems and Imaging*, 1, 399–422.
- Nir, T., Bruckstein, A. M., & Kimmel, R. (2008). Over-parameterized variational optical flow. *International Journal of Computer Vision*, 76(2), 205–216.
- Niu, Y., Feng, W., & Liu, F. (2012). Enabling warping on stereoscopic images. *ACM Transactions on Graphics*, 31(6), 183.
- Press, W. H., Vetterling, W. T., Teukolsky, S. A., & Flannery, B. P. (2002). *Numerical recipes in C++: The art of scientific computing*. New York: Cambridge University Press.
- Ren, X. (2008). Local grouping for optical flow. In *IEEE International Conference on Computer Vision and Pattern Recognition*.
- Rudin, L. I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1–4), 259–268.
- Sand, P., & Teller, S. (2008). Particle video: Long-range motion estimation using point trajectories. *International Journal of Computer Vision*, 80(1), 72–91.
- Schmidt, U., Gao, Q., & Roth, S. (2010). A generative perspective on MRFs in low-level vision. In *IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 1751–1758).
- Shulman, D., & Herve, J.-Y. (1989). Regularization of discontinuous flow fields. In *Workshop on Visual Motion* (pp. 81–86).
- Steinbrücker, F., Pock, T., & Cremers, D. (2009). Large displacement optical flow computation without warping. In *IEEE International Conference on Computer Vision* (pp. 1609–1614).
- Sun, D., Roth, S., & Black, M. J. (2010a). Secrets of optical flow estimation and their principles. In *IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 2432–2439).
- Sun, D., Roth, S., Lewis, J. P., & Black, M. J. (2008). Learning optical flow. In *European Conference on Computer Vision* (Vol. III, pp. 83–97).
- Sun, D., Sudderth, E. B., & Black, M. J. (2010b). Layered image motion with explicit occlusions, temporal consistency, and depth ordering. In *Advances in Neural Information Processing Systems* (pp. 2226–2234). Cambridge, MA: MIT Press.
- Sun, D., Sudderth, E. B., & Black, M. J. (2012). Layered segmentation and optical flow estimation over time. In *IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 1768–1775).
- Sun, D., Wulff, J., Sudderth, E. B., Pfister, H., & Black, M. J. (2013). A fully-connected layered model of foreground and background flow. In *IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 1768–1775).
- Szeliski, R. (2010). *Computer vision: Algorithms and applications*. New York: Springer.
- Vaudrey, T., & Klette, R. (2009). Residual images remove illumination artifacts! In *Pattern Recognition (Proceedings of DAGM)* (pp. 472–481). Berlin: Springer.
- Wang, J. Y. A., & Adelson, E. H. (1994). Representing moving images with layers. *IEEE Transactions on Image Processing*, 3(5), 625–638.
- Wedel, A., Pock, T., Braun, J., Franke, U., & Cremers, D. (2008a). Duality TV-L1 flow with fundamental matrix prior. In *Image and Vision Computing New Zealand*.
- Wedel, A., Pock, T., Zach, C., Cremers, D., & Bischof, H. (2008b). An improved algorithm for TV-L1 optical flow. In *Dagstuhl Motion Workshop* (pp. 23–45).
- Wedel, A., Pock, T., & Cremers, D. (2009). Structure- and motion-adaptive regularization for high accuracy optic flow. In *IEEE International Conference on Computer Vision* (pp. 1663–1668).
- Werlberger, M., Pock, T., & Bischof, H. (2010). Motion estimation with non-local total variation regularization. In *IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 2464–2471).
- Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., & Bischof, H. (2009). Anisotropic Huber-L1 optical flow. In *Proceedings of the British Machine Vision Conference* (pp. 108.1–108.11).
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83.
- Xiao, J., Cheng, H., Sawhney, H., Rao, C., & Isnardi, M. (2006). Bilateral filtering-based optical flow estimation with occlusion detection. In *European Conference on Computer Vision* (Vol. I, pp. 211–224).

- Xu, L., Chen, J., & Jia, J. (2008). A segmentation based variational model for accurate optical flow estimation. In *European Conference on Computer Vision* (Vol. I, pp. 671–684).
- Xu, L., Jia, J., & Matsushita, Y. (2012). Motion detail preserving optical flow estimation. *IEEE Transaction on Pattern Analysis Machine Intelligence*, 34(9), 1744–1757.
- Yoon, K., & Kweon, I. (2006). Adaptive support-weight approach for correspondence search. *IEEE Transaction on Pattern Analysis Machine Intelligence*, 28(4), 650–656.
- Zach, C., Pock, T. & Bischof, H. (2007). A duality based approach for realtime TV-L1 optical flow. In *Pattern Recognition (Proceedings of DAGM)* (pp. 214–223).
- Zimmer, H., Bruhn, A., & Weickert, J. (2011). Optic flow in harmony. *International Journal of Computer Vision*, 93(3), 368–388.
- Zimmer, H., Bruhn, A., Weickert, J., Valgaerts, L., Salgado, A., Rosenhahn, B. & Seidel, H.-P. (2009). Complementary optic flow. In *Energy Minimization Methods in Computer Vision and Pattern Recognition* (pp. 207–220).
- Zitnick, C., Jojic, N., & Kang, S. B. (2005). Consistent segmentation for optical flow estimation. In *IEEE International Conference on Computer Vision* (Vol. 2, pp. 1308–1315).