A General Dense Image Matching Framework Combining Direct and Feature-based Costs

Jim Braux-Zin[1,2]
jim.braux-zin@cea.fr

Romain Dupont[1]
romain.dupont@cea.fr

Adrien Bartoli[2]
adrien.bartoli@gmail.com

[1] CEA, LIST, France

[2] ISIT, Université d'Auvergne/CNRS, France

Abstract

Dense motion field estimation (typically optical flow, stereo disparity and surface registration) is a key computer vision problem. Many solutions have been proposed to compute small or large displacements, narrow or wide baseline stereo disparity, but a unified methodology is still lacking. We here introduce a general framework that robustly combines direct and feature-based matching. The feature-based cost is built around a novel robust distance function that handles keypoints and "weak" features such as segments. It allows us to use putative feature matches which may contain mismatches to guide dense motion estimation out of local minima. Our framework uses a robust direct data term (AD-Census). It is implemented with a powerful second order Total Generalized Variation regularization with external and self-occlusion reasoning. Our framework achieves state of the art performance in several cases (standard optical flow benchmarks, wide-baseline stereo and non-rigid surface registration). Our framework has a modular design that customizes to specific application needs.

Introduction

A dense motion field, also called optical flow, is a very useful cue for problems such as tracking, segmentation, localization and reconstruction, or non-rigid surfaces registration. Optical flow estimation is an old computer vision problem. While early techniques were patch-based [19], current ones estimate dense flow fields with variational methods built upon the work by Horn and Schunk [16] by coupled minimization of a data term – often based on the brightness constancy assumption – and regularization. The use of non-quadratic norms such as L^1 or Huber [30] and illumination-invariant data terms [35, 20, 25] has led to increasingly accurate and robust algorithms. Occlusions can be somehow handled with anisotropic regularizers [30, 36] but those are very application-dependent and degrade the output if not well tuned. Current best results are achieved when occlusions are explicitly modeled [32, 13]. Coarse-to-fine warping improves global convergence by making the assumption that the motion of smaller structures is similar to the motion of larger structures. These techniques may fail in the following two cases. First in the presence of image structures whose size is smaller than their relative displacement magnitude ; second, in the wide-baseline or non-rigid settings where the images can be too different at coarser resolutions to facilitate an accurate initialization.

Contrary to those methods, feature detection and description aims at obtaining sparse global matches. Most point descriptors such as SIFT [18] and SURF [3] use the concept of histograms of gradient. Another kind of common features is segments [28, 29]. Wang *et al*. proposed a technique [28] for wide-baseline segment matching which encodes semi-local geometric information about the scene and is thus a good fit for low-texture or high perspective distortion image pairs. The amount of outliers depends on the quality of the feature, detector and descriptor used but there is no way to suppress them without strong assumptions on the scene.

The research on non-rigid surface registration offers a good example of the duality of features and whole image information. Surface tracking – frame to frame estimation of a deformation – offers the best accuracy when using a direct pixel-wise cost [13, 12] but those methods, similarly to standard optical flow techniques are limited to small updates at each frame. On the other hand there is the problem of surface detection: estimation of the deformation from only one frame and the flat template image. This is usually done from feature matches which are first filtered to remove outliers and then fitted by a warp such as a Thin-Plate Spline [23]. Recent works [24] aim at initializing an image-based method from a feature-based warp but the two steps are still independent and if the warp is too inaccurate in some areas, the final step is unlikely to recover the true deformation. This duality is also present in 3D reconstruction where narrow-baseline stereo is solved with dense variational methods [25] but wide-baseline reconstruction is usually based on features such as in keyframe-based Simultaneous Localization

And Mapping [21] (we do not consider the multi-view case, which can be seen as several narrow-baseline problems). Dense descriptors [27, 17] tend to blur the line between image and feature-based costs but are still sparsely used and need to be embedded in a costly discrete global optimization.

Surprisingly few attempts have been made to use image and feature informations at the same time. Wills *et al.* [31] propose a RANSAC approach to dense motion segmentation and estimation from feature matches. The layered approach yields an easy understanding of the scene motions but is restricted in accuracy by the deformation model (planar and non-rigid models were proposed). Xu *et al.* [32] feed a discrete optimization step with candidate displacement vectors from, among other, SIFT features and patch matches. The output is then refined with a variational regularization. This gives very accurate results for small or large displacements but the process is computationally expensive. Brox *et al.* [9] inspired our work with a successful approach coupling descriptor matching with variational optimization in the same process. Their method remains a reference in optical flow estimation for its robustness and accuracy, but we identified several limitations. It only uses custom descriptors computed densely or on a fixed spatial grid and the tight coupling between the feature matching and the cost function (see Section 1) prevents from using state of the art features. As we will see, this puts unnecessary restrictions on the scope of the method. Moreover, the resolution process is also specific to the method and does not use current faster primal-dual based algorithms.

We propose a novel approach to widen the range of applicability of dense variational methods. This is achieved by combining direct and feature-based costs in one process with a focus on flexibility and robustness to easily exploit any descriptor and detector. Feature detection and matching is done only once on the full-resolution image pair but the feature matches guide the optical flow out of local minima during the whole optimization. After a brief introduction to the LDOF [9] method in Section 1, we introduce our framework built around the second-order Total Generalized Variation [8] for regularization in Section 2.1, a robust direct cost with occlusions handling in Section 2.2, and a novel feature-based cost able to handle unfiltered matches of point features or weakly localized segments in Section 2.3. Each of these building blocks can easily be upgraded to take advantage of future work on regularization, image-based cost or feature matching.

Notation. We consider a pair of images I_0 and I_1. The motion field is estimated over the image domain Ω of I_0. The images are modeled as continuous real-valued functions by interpolating the pixel intensities. The L^2 Euclidean norm is noted $\|.\|$ and the L^1 norm is noted $|.|$. Vectors and vector-valued functions are noted in bold lowercase ($\mathbf{x}$) and matrices in bold uppercase ($\mathbf{J}$). The estimated motion field is called $\mathbf{u} : \Omega \to \mathbb{R}^2$ such as for all $\mathbf{x} \in \Omega$, $I_0(\mathbf{x}) \approx I_1(\mathbf{x} + \mathbf{u}(\mathbf{x}))$.

1. Large Displacement Optical Flow

In this section we summarize the work of Brox et al. [9] who proposed the first introduction of rich descriptors in variational optical flow computation through the addition of a new term in the cost function:

$$\begin{aligned} C_{\text{LDOF}}(\mathbf{u}) = \; & C_{\text{color}}(\mathbf{u}) + \gamma\, C_{\text{gradient}}(\mathbf{u}) && (1) \\ & + \alpha\, C_{\text{smooth}}(\mathbf{u}) && (2) \\ & + \beta \sum_{i=1}^{N} C_{\text{match}}(\mathbf{u}, \mathbf{u}_{\text{match}}^{(i)}) + \underbrace{C_{\text{desc}}(\mathbf{u}_{\text{match}}^{(i)})}_{\text{nearest neighbors}} && (3) \end{aligned}$$

with (1) the direct data term, (2) the regularizer and (3) the feature integration cost. $C_{\text{match}}(\mathbf{u}, \mathbf{u}_{\text{match}}^{(i)})$ is defined as

$$\int_\Omega \delta_i(\mathbf{x}) \rho_i(\mathbf{u}_{\text{match}}^{(i)}) \Psi(\|\mathbf{u}(x) - \mathbf{u}_{\text{match}}^{(i)}\|)\, \mathrm{d}\mathbf{x} \qquad (4)$$

where for each feature match $i \in 1 \dots N$, $\mathbf{u}_{\text{match}}^{(i)}$ is the displacement induced by the match, δ_i is an indicator function for the affected pixels and $\rho_i(\mathbf{u}_{\text{match}}^{(i)})$ the quality of the match. $\Psi : x^2 \to \sqrt{x^2 + \epsilon}$ is a robust cost function approximating the L^1 norm. Brox *et al.* proposed two types of features to associate *a priori* displacements to affected pixels: regions using SIFT [18] and color based descriptors, and points using Histogram of Oriented Gradients [11] or Geometric Blur [4] over a fine regular grid. Nearest neighbors in feature space are selected as potential matches ; this processus is represented by the term $C_{\text{desc}}(\mathbf{u}_{\text{match}})$ which does not intervene in the variational optimization. The direct cost term (1) consists in the combination of brightness and gradient differences while the regularization cost (2) is an approximation of the Total Variation, both with the same robust norm Ψ. The resolution uses variable decoupling and alternated standard primal minimization with linearized Lagrange equations in a coarse-to-fine warping scheme.

Our framework does not share many implementation details but was inspired by two main findings. First, Brox *et al.* showed that while not increasing the best case accuracy, feature matches help drive optical flow estimation out of local minima when the common assumptions of coarse-to-fine warping are violated. Second, they highlighted the nice behavior of coarse-to-fine warping naturally giving a decreasing weight to features as they go from quasi-dense at coarse levels to sparser at fine levels.

While it leads to impressive results, we find this model to be overly restrictive. Equation (4) needs the features to be well localized (to translate matches to displacements) and to provide a match quality score to minimize the impact of outliers. This limits the scope of suitable features and introduces a tight coupling between the feature matcher and the variational solver. The logic behind this choice was to let the variational solver "do the matching" by selecting the best feature amongst the neighbors in feature space. However the conclusion in [9] is that incorporating only the best neighbor gives better results by getting rid of conflicting costs. We propose a generalized approach that relaxes those constraints for an easy plugging in of current and future feature matchers, see Section 2.3.

2. Proposed Framework

Our cost function has the following form:

$$C_{\text{ours}}(\mathbf{u}) = \lambda\, C_{\text{direct}}(\mathbf{u}) + \text{TGV}^2(\mathbf{u}, \alpha_0, \alpha_1) + \beta \sum_{i=1}^{N} C_{\text{match}}(\mathbf{u}, \mathcal{F}_i). \tag{5}$$

We use a variational model built around the recently introduced second order Total Generalized Variation [8] (TGV^2) regularization which favors piecewise-affine displacement fields as detailed in Section 2.1. Several direct data terms C_{direct} can be used, we list three in Section 2.2. Contrary to LDOF in (4), C_{match} is here a function of the distance of $\mathbf{x} + \mathbf{u}(\mathbf{x})$ to the target feature and not the difference between $\mathbf{u}(\mathbf{x})$ and an *a priori* displacement. This distinction allows us to incorporate features which are not "fully localized" such as segments. The feature-based cost function is explained Section 2.3.

2.1. Second Order Total Generalized Variation

Definition. The two components u_x and u_y of the optical flow are independently regularized. In this section we use the notation u to design any of those two scalar fields. The Total Variation $\text{TV}(u) = \int_\Omega |\nabla u(\mathbf{x})| \, \mathrm{d}\mathbf{x}$ has been one of the most used regularizers in the literature thanks to its nice edge-preserving behaviour and the development of efficient algorithms [10]. It suffers however from staircasing effects in case of smooth flows. Several alternatives have been proposed such as the Huber-L^1 regularization [30] but we focus here on the elegant Total Generalized Variation [8] which brings the discontinuity preserving quality of the Total Variation to arbitrarily high order derivatives. It is defined in the Legendre-Fenchel dual space by:

$$\text{TGV}^k_\alpha(u) = \sup_{\mathbf{v}} \left\{ \int_\Omega u(\mathbf{x}) \, \text{div}^k \, \mathbf{v}(\mathbf{x}) \,\middle|\, \mathbf{v} \in \mathcal{C}^k_c(\Omega, \mathbb{R}^d), \; \| \text{div}^l \, \mathbf{v} \|_\infty \leq \alpha_l, \quad l = 0, \dots, k-1 \right\}$$

where $\mathcal{C}^k_c$ is the space of order k tensors, α_l are tunable weights and div^k is the k-divergence for symmetric tensor fields formally defined in [8]. The primal formulations make the link with the Total Variation. Most notably we have:

$$\text{TGV}^1_\alpha(u) = \alpha \text{TV}(u) = \alpha \int_\Omega |\nabla u(\mathbf{x})| \, \mathrm{d}\mathbf{x}$$

$$\text{TGV}^2_\alpha(u) = \min_{\substack{\mathbf{w} \in \\ \Omega \to \mathbb{R}^2}} \left\{ \alpha_1 \int_\Omega |\nabla u - \mathbf{w}| \, \mathrm{d}\mathbf{x} + \alpha_0 \int_\Omega |\nabla \mathbf{w}| \, \mathrm{d}\mathbf{x} \right\}.$$

TGV^1 favors piecewise constant solutions while TGV^2 favors piecewise affine solutions, more desirable for stereo and optical flow estimations.

Primal-dual TGV^2 optimization. The efficient Chambolle & Pock primal-dual optimization scheme [10] can be used to optimize the discretized TGV^2 model. We refer the reader to [7, 25] for a detailed introduction to the algorithm with practical details such as the discretized operators. A standard coarse-to-fine warping with a subsampling factor $s \in [0.5, 1[$ embeds those iterations to avoid local minima and increases the convergence rate.

2.2. Direct Data Term

Pixel-wise costs. Our warping scheme is based on the iterative linearization of the cost function. Any sufficiently smooth pixel-wise data term can be used, but complexity constraints must be taken into account to keep reasonable performances as the direct cost is going to be evaluated hundreds of times at each pixel. Linear interpolation allows us to see the discrete pixel intensities as a continuous smooth function over the image domain. In the following we derive the direct costs used in our experiments.

The Absolute Difference cost is based on the common brightness constancy assumption. It is the simplest, most used term. It is robust to image deformations but degrades quickly in the presence of illumination changes:

$$C_{\text{AD}}(\mathbf{x}, \mathbf{u}) = |I_0(\mathbf{x}) - I_1(\mathbf{x} + \mathbf{u}(\mathbf{x}))|. \tag{6}$$

We use the ternary Census [35] transform as explained in [25]. The Census transform encodes local structure in a fixed size window and provides a matching robust to additive or multiplicative illumination changes. It looses however some localization accuracy because of the underway discretization of pixel differences. A bigger window makes the transform more discriminative but less robust to distortions:

$$C_{\text{Census}}(\mathbf{x}, \mathbf{u}) = \Delta(C(I_0, \mathbf{x}), C(I_1, \mathbf{x} + \mathbf{u}(\mathbf{x}))), \tag{7}$$

where Δ is the Hamming distance and $C(I, \mathbf{x})$ is the Census transform of the image I at pixel $\mathbf{x}$. The AD-Census [20] approach combines the previous ones to increase the accuracy of Census and preserves its robustness:

$$C_{\text{ADC}}(\mathbf{x}, \mathbf{u}) = 2 - \exp\left(-\frac{C_{\text{AD}}}{\mu_0}\right) - \exp\left(-\frac{C_{\text{Census}}}{\mu_1}\right). \tag{8}$$

Occlusions handling. For non-trivial motion estimation, occlusions should be taken into account. Following [13] we distinguish two kinds of occlusions: external and self-occlusions. External occlusions are handled by clipping the data term to a threshold to prevent outliers from disturbing the whole estimation. This threshold depends on the application and on the data term used. We found in the scenarios tested with the AD-Census data-term that a value of 50% of the maximum direct cost value significantly increases robustness without deteriorating results in the absence of occlusions. Self-occlusions appear when a rigid scene is observed from different viewpoints or when a deformable surface folds. They can be handled gracefully with the theoretically justified assumption that the derivative of the warp then vanishes in one direction [13]. Given the warp $\mathcal{W}(\mathbf{x}) = \mathbf{x} + \mathbf{u}(\mathbf{x})$, $\mathbf{x}$ is occluded if:

$$\exists \mathbf{d} \mid \|\mathbf{d}\| = 1 \text{ and } \nabla_{\mathbf{d}} \mathcal{W}(\mathbf{x}) = 0 \text{ i.e. } \nabla_{\mathbf{d}} \mathbf{u}(\mathbf{x}) = -\mathbf{d} \tag{9}$$

where $\nabla_{\mathbf{d}} \mathbf{u}(\mathbf{x}) \approx \frac{\mathbf{u}(\mathbf{x}+\epsilon\mathbf{d}) - \mathbf{u}(\mathbf{x}-\epsilon\mathbf{d})}{2\epsilon}$ is the finite central differences based partial derivative in direction $\mathbf{d}$. It has been shown [13] that the smallest squared partial derivative σ_0 is linked to the Jacobian $\mathbf{J}$ of the warp $\mathcal{W}$ by $\sigma_0 = \min_{\|\mathbf{d}\|=1} \mathbf{d}^{\mathrm{T}} \mathbf{J}^{\mathrm{T}} \mathbf{J} \mathbf{d}$ and after spectral decomposition of $\mathbf{O} = \mathbf{J}^{\mathrm{T}} \mathbf{J}$:

$$\sigma_0 = \frac{1}{2} \left(\mathbf{O}_{11} + \mathbf{O}_{12} - \sqrt{(\mathbf{O}_{11} - \mathbf{O}_{22})^2 + 4\mathbf{O}_{12}^2} \right). \tag{10}$$

A smooth step function $\mathcal{S}(x, k, r) = \frac{1}{1+\exp(-k(x-r))}$ then translates σ_0 to an occlusion probability:

$$\mathcal{P}_{\text{occ}} = 1 - \mathcal{S}(\sigma_0, 40, 0.1). \tag{11}$$

The direct data term is not to be trusted on occluded areas so we multiply it by $1 - \mathcal{P}_{\text{occ}}$ before including it in the global cost function (5).

2.3. Feature-based Data Term

As hinted in Section 1, we draw inspiration from the LDOF feature based cost (4) with relaxed constraints to propose a more general approach:

$$C_{\text{match}}(\mathbf{u}, \mathcal{F}_i) = \int_{\Omega} \rho_i(\mathbf{x}) D(\mathbf{x}, \mathbf{u}(\mathbf{x}), \mathcal{F}_i) \, \mathrm{d}\mathbf{x} \tag{12}$$

$$\begin{aligned} D(\mathbf{x}, \mathbf{u}(\mathbf{x}), \mathcal{F}_i) = {} & c\Gamma_\sigma[D_{\text{f}}(\mathbf{x} + \mathbf{u}(\mathbf{x}), \mathcal{F}_i)] \\ & + (1-c)\, \Gamma_\sigma[D_{\text{ap}}(\mathbf{x}, \mathbf{u}(\mathbf{x}), \mathcal{F}_i)], \end{aligned} \tag{13}$$

where ρ_i is the influence function of the feature, Γ_σ is a robust estimator, D_{f} is the main *feature* distance, D_{ap} the *a priori match* distance for weakly localized features and c a weighting factor. Explanations follow.

Influence function. Our framework is not restricted to a regular grid of descriptors, and most features have a sub-pixel accurate location. The influence function translates this property by the use of linear interpolation. Given a point feature i located at $\mathbf{x}_f = \mathbf{x}_{f_0} + \mathbf{dx}$, $\mathbf{x}_{f_0} = \text{floor}(\mathbf{x}_f)$, $\overline{\mathbf{dx}} = (1,1)^T - \mathbf{dx}$, its influence function ρ_i is defined for the four affected pixels as:

$$\begin{aligned} \rho_i\left(\mathbf{x}_{f_0}\right) = \overline{\mathbf{dx}}_x \overline{\mathbf{dx}}_y \,|\, \rho_i\left(\mathbf{x}_{f_0} + (0,1)^T\right) = \overline{\mathbf{dx}}_x \mathbf{dx} \\ \rho_i\left(\mathbf{x}_{f_0} + (1,0)^T\right) = \mathbf{dx}_x \overline{\mathbf{dx}}_y \,|\, \rho_i\left(\mathbf{x}_{f_0} + (1,1)^T\right) = \mathbf{dx}_x \mathbf{dx}_y. \end{aligned}$$

Segment features receive the same treatment by considering them as a dense set of pixels. The influence function is then an anti-aliased discrete representation of the segment.

Robust cost function. Contrary to [9], we do not rely on any exterior match quality measure, often unreliable or unavailable. We consider that a match is an inlier if it is compatible with the overall motion. To this end, we use the non-convex Geman McClure M-estimator $\Gamma_\sigma(x) = \frac{x^2}{\sigma + x^2}$ whose influence function $\frac{\mathrm{d}\Gamma_\sigma(x)}{\mathrm{d}x} \propto \frac{x}{(\sigma + x^2)^2}$ tends toward zero when x grows large. We use a small $\sigma = 0.2$ for an efficient implicit outliers filtering. It can be seen that at coarse levels, where several features affect each pixel, a vote takes place where the outvoted matches durably loose influence. At finer level, the regularization and the direct data term influences are more important and should converge to the desired optimum.

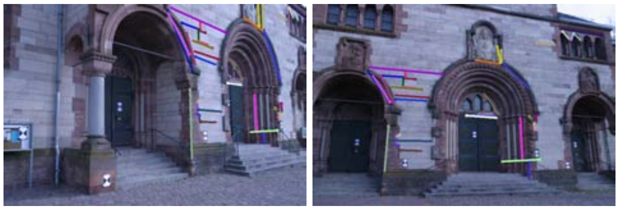
Feature distances. The feature distances are the standard cost associated to features. Here follows definitions for point distance and line distance.

Point features are a mature and active field of research. A wide choice of descriptors provides an optimal trade-off between speed (SURF [3]) and robustness (ASIFT [34]). Using corner detectors [15, 26] allows for more accurately localized features compared to the regular grid used in LDOF [9]. The feature distance of points is the Euclidean distance. For a point feature $\mathcal{F}_i = \mathbf{x}_i$:

$$D_{\text{f}}^{(\text{point})}(\mathbf{x}, \mathcal{F}_i) = \|\mathbf{x} - \mathbf{x}_i\|. \tag{14}$$

As demonstrated in [9], the influence of point features naturally decreases during the coarse-to-fine warping. Indeed the influence of features is related to the area they occupy. At each upsampling step with a factor $s > 1$, the relative area covered by a pixel is multiplied by $s^{-2} < 1$. As we will see, *a priori matches* are only needed for weakly-localized features so $c = 1$ for point features.

The segment matching algorithm proposed by Wang [28] is interesting because it does not rely on photometric similarities but encodes semi-global structure and is robust to wide-baseline perspective distortion. A matching example



(a) I_0 (b) I_1

Figure 1. Example of segment matches using [28]. Best viewed in color.

is displayed in Figure 1. Segment matches lie on the same line but their endpoints are not guaranteed to be matched in both images ; in fact, due to occlusions or different image boundaries it is rarely the case. This means that the proper distance feature to use is the orthogonal distance to line, which constrains only one dimension. Given a segment feature defined by its endpoints $\mathcal{F}_i = (\mathbf{x}_{i_b}, \mathbf{x}_{i_e})$:

$$D_{\text{f}}^{(\text{segment})}(\mathbf{x}, \mathcal{F}_i) = \frac{\|(\mathbf{x}_{i_e} - \mathbf{x}_{i_b}) \times (\mathbf{x} - \mathbf{x}_{i_b})\|}{\|\mathbf{x}_{i_e} - \mathbf{x}_{i_b}\|}. \tag{15}$$

During the downsampling for coarse-to-fine processing, the area of the image affected by segments only shrinks in one dimension. To make the influence of segment features vanish at the same rate as point features, the influence functions ρ_i are multiplied by s^{-1} at each upsampling step.

A priori matches. Even though the segment matches constrain only one dimension, the remaining degree of freedom cannot be left fully unconstrained. We introduce the concept of *a priori matches* which makes the assumption of a linear mapping between segment matches. Given a segment $\mathcal{F}_i^{(0)} = (\mathbf{x}_{i_b}^{(0)}, \mathbf{x}_{i_e}^{(0)})$ in I_0 and its match $\mathcal{F}_i^{(1)} = (\mathbf{x}_{i_b}^{(1)}, \mathbf{x}_{i_e}^{(1)})$ in I_1, the *a priori match* $\mathbf{x}_{\text{ap}}^{(1)}$ of the point $\mathbf{x}^{(0)} \in \mathcal{F}_i^{(0)}$ is defined by:

$$t = \frac{\left\langle (\mathbf{x}^{(0)} - \mathbf{x}_{i_b}^{(0)}), (\mathbf{x}_{i_e}^{(0)} - \mathbf{x}_{i_b}^{(0)}) \right\rangle}{\|\mathbf{x}_{i_e}^{(0)} - \mathbf{x}_{i_b}^{(0)}\|^2}$$

$$\mathbf{x}_{\text{ap}}^{(1)} = \mathbf{x}_{i_b}^{(1)} + t \cdot (\mathbf{x}_{i_e}^{(1)} - \mathbf{x}_{i_b}^{(1)})$$

and the corresponding distance is defined by:

$$D_{\text{ap}}^{(\text{segment})}(\mathbf{x}, \mathbf{u}(\mathbf{x}), \mathcal{F}_i) = \|\mathbf{x} + \mathbf{u}(\mathbf{x}) - \mathbf{x}_{\text{ap}}^{(1)}\|. \tag{16}$$

This hypothesis is only true for fronto-parallel segments with perfectly matched endpoints. However it is most of the time not far from the truth and can be used as a cue to guide the optical flow estimation at coarse levels. As shown in Table 1, *a priori matches* are not an accurate prior and degrade a lot the results if used as the sole constraint in either the small or wide-baseline settings. However, with a more balanced coefficient $c = 0.5$ the results are improved in the small-baseline case compared to using only the feature distance without loss of accuracy in the wide-baseline case. This approach could be easily extended to other weakly localized features like regions or contours.

c		Narrow-baseline	Wide-baseline
0	only *a priori matches*	92.6%	0%
0.5		93.9%	41.7%
1	no *a priori matches*	93.3%	41.7%

Table 1. Influence of the inclusion of *a priori matches* on the proportion of inlier depths values (error smaller than 5% of the depth range) when using segment features. The narrow-baseline results are obtained with the image pair $1-2$ from the `herzjesu` benchmark and the wide-baseline with the image pair $6-1$. Details in Section 2.3.

3. Experimental Results

In this section we demonstrate the validity and the versatility of our approach on several benchmarks. We start by the standard optical flow benchmarks and then show promising results on wide-baseline stereo and non-rigid surface registration. When not stated otherwise, we use the following parameters set: $\lambda = 6$, $\beta = 0.5$, $\alpha_0 = 4$, $\alpha_1 = 1$, 20 warp of 40 iterations each and a subsampling factor $s = 0.8$. The direct data term is AD-Census with a 3×3 window, $\mu_0 = 1$ and $\mu_1 = 0.25$. For keypoint matches, we use the FAST detector [26] and the SIFT descriptor [18] from the OpenCV library [6], all with default parameters for easy reproduction. A simple cross-check filter removes the obvious ambiguous matches.

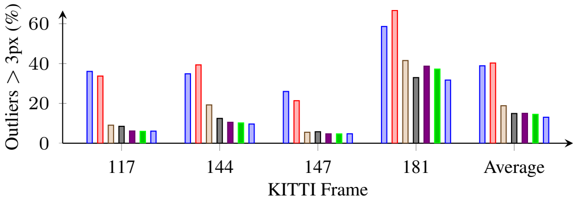
3.1. Small-Baseline

We speak of small-baseline setting when the overall motion magnitude is low. This is typically the case in frame-to-frame optical flow benchmarks. We evaluate our method on two of them.

KITTI benchmark. The KITTI [14] benchmark is a recent benchmark composed of real-world images captured from a moving vehicle. They present challenges such as specularities, dominant non-fronto-parallel surfaces (like the road), high variability in the displacement magnitudes and in the illumination conditions. At the time of writing, our method is the top ranked true 2D optical flow method in this benchmark as can be seen in Table 2. It means that even for relatively small displacements (frame-to-frame optical flow) our method outperforms state of the art. To analyze the impact of our contributions on this benchmark, we use the challenging "large displacements" selection of frames from the Special Session on Robust Optical Flow [1]. In Figure 2 we compare our method to LDOF [9] that inspired our framework, TGV2CENSUS [25] based on the same Total Generalized Variation regularization and MDPOF [32], the

R	Method	Out-Noc	Out-All	Avg-Noc	Avg-All	Time
1	PR-Sf+E	4.08 %	7.79 %	0.9 px	1.7 px	200 s
2	PCBP-Flow[ms]	4.08 %	8.70 %	0.9 px	2.2 px	3 min
3	MotionSLIC[ms]	4.36 %	10.91 %	1.0 px	2.7 px	11 s
4	PR-Sceneflow	4.48 %	8.98 %	1.3 px	3.3 px	150 s
5	TGV2ADCSIFT	6.55 %	15.35 %	1.6 px	4.5 px	12 s (GPU)
6	Data-Flow	8.22 %	15.78 %	2.3 px	5.7 px	3 min
7	fSGM	11.03 %	22.90 %	3.2 px	12.2 px	60 s
8	TGV2CENSUS	11.14 %	18.42 %	2.9 px	6.6 px	4 s (GPU)

Table 2. Results on the KITTI benchmark [14]. Methods 1 and 4 (anonymous) are scene-flow based and unpublished but probably not comparable. Methods 2 and 3 [33] are one-dimensional motion stereo estimation methods. Our method ranked 5 is the top true 2D optical flow method.



	Method	Avg. error
🟦	Large Displacement Optical Flow [9]	38.88%
🟥	Motion Detail Preserving Optical Flow [32]	40.28%
🟫	TGV2CENSUS [25] ($\lambda = 5$, $\alpha_0 = 4$, Census, 30 warps)	18.83%
⬛	$\lambda = 6$, $\alpha_0 = 1$ with Census, 20 warps	14.92%
🟪	$\lambda = 6$, $\alpha_0 = 1$ with AD-Census, 20 warps	15.02%
🟩	…and self-occlusions handling	14.52%
🟦	…and incorporating FAST-SIFT matches	13.04%

Figure 2. Quantitative evaluation on the “large displacements” selection of KITTI image pairs [1].

only competitive direct and feature-based method. We see that the affine regularization of TGV^2 is the single most important parameter. Then, the biggest gain is achieved thanks to our parameter set: the more relaxed α_0 weight creates less smooth displacement fields. The AD-Census improvement is not visible on the average error because of the outlier at frame 181. Even if the small-baseline setting is not the most needing of self-occlusions handling and incorporation of keypoint matches, those two additions bring meaningful gains. The OpenCV FAST-SIFT matcher produced for each pair approximately 4000 matches in 4 seconds on a 6×2.40GHz Intel Xeon CPU and the variational estimation took 8s on average with an outdated NVidia GeForce GTX 460 GPU. The focus on this paper is on producing competitive output with state of the art in various settings and not on performance. We are convinced that with better hardware, optimization and parameter tweaks an execution time inferior to one second is reachable.

Middlebury benchmark. The venerable Middlebury benchmark [2] has been a reference for years. However, the proposed image pairs, either synthetic or from a controlled environment are not representative of real-world motions. We think it is interesting anyway to show that the presence of outliers in matches do not noticeably degrade the results. We use unfiltered SURF [3] matches for this experiment. They are of very low quality as can be seen in Figure 3b. The TGV^2 regularization is not adequate for the controlled motions of the Middlebury benchmark, mostly piecewise constant, so we use a standard TV regularizer in this experiment by setting $\alpha_0 = \infty$. The results are displayed in Table 3a. The great resilience of our algorithm to mismatches comes from the use of a non-convex robust estimator which penalizes matches which are not coherent with others.

3.2. Wide-Baseline Stereo

The increased robustness of our approach allows us to explore applications such as wide-baseline stereo, previously unreachable to optical flow techniques. Tola et al. [27] published interesting datasets along with impressive results using an innovative robust dense descriptor and Graph-Cut based discrete optimization [5]. We ran our algorithm on the `herzjesu` dataset with segment matches [28] and a feature weight $\gamma = 5$ while constraining the displacement vectors on the epipolar lines. We adopted the same image matrix form than in their paper for easy comparison of the results in Figure 3. For extreme perspective distortions, the AD-Census data term shows its limits but we see that the segment features greatly increase the convergence basin and allows for comparable results on most image pairs. Moreover, our continuous variational approach is faster than the Graph-Cut ones and is not restricted to one dimension. One can also note that our occlusion handler, although coming from the deformable surface field [13] is also suitable for rigid settings.

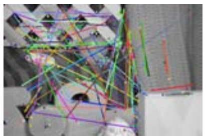
3.3. Deformable Surface Detection

Another interesting field of application where optical flow methods have so far been unadapted is the detection of deformable surfaces. Given a flat surface template and an image of this same surface with non-rigid deformation, the problem is to estimate the pixel correspondences between the two images. For non-trivial deformations, pixel-based direct methods need an initialization close to the solution [24] and are mostly used on video sequences (surface registration). Non-rigid surface detection methods are feature-based and usually adopt a two step approach: features filtering and fitting of a warp (Thin-Plate Spline or Free-Form Deformation).

To show the suitability of our method to deformable surface detection and produce some quantitative results, we generate a synthetic deformation using the Matlab toolbox from [22]. We obtain point matches using the SIFT detector and descriptor. We compare ourself with the feature-based method [24] which is used to filter the matches and then fit a standard Free-Form Deformation warp. We also add the LDOF optical flow method as the most robust optical flow method in the state of the art. We show in Figure 4 that

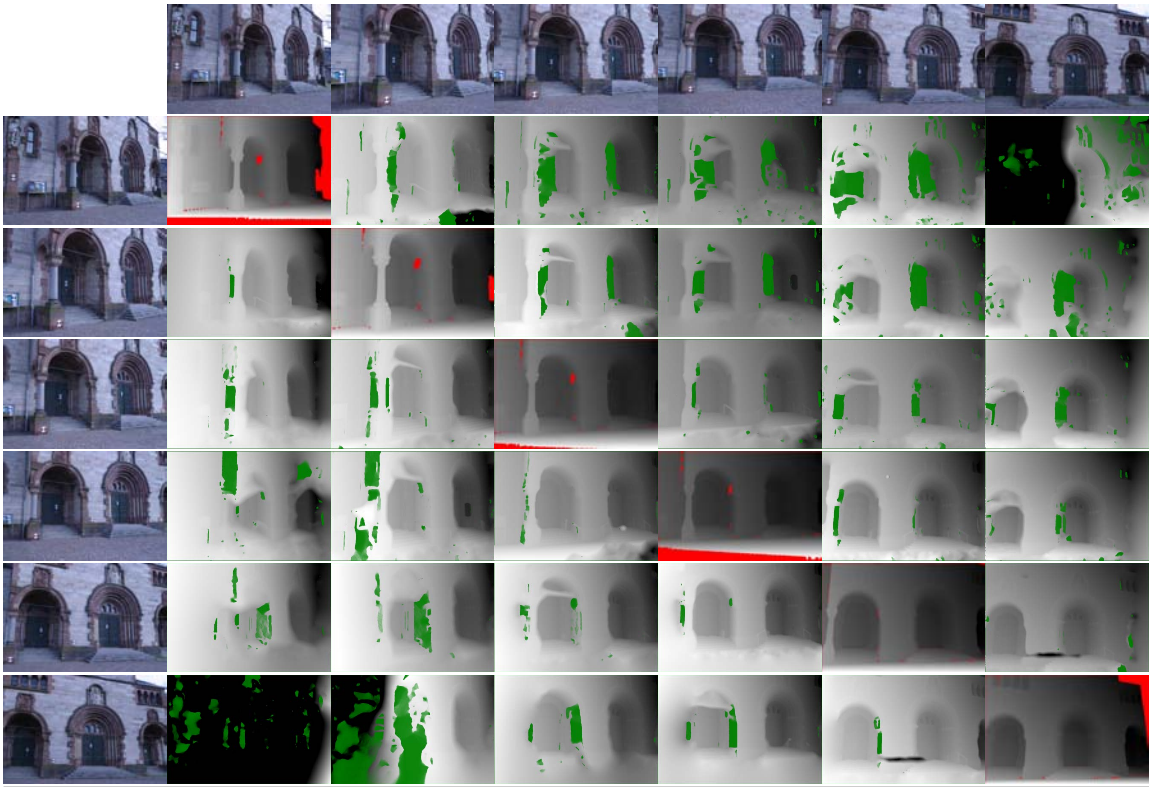
Method	Dimetrodon	Grove2	Grove3	Hydrangea	RubberWhale	Urban2	Urban3	Venus	Average
LDOF [9]	0.12	0.18	0.70	0.18	0.13	0.38	0.82	0.38	0.36
Ours	0.12	0.18	0.71	0.18	0.13	0.46	0.60	0.26	0.33
Ours with matches	0.13	0.18	0.72	0.18	0.13	0.45	0.59	0.26	0.33

(a)



(b)

Table 3. Benchmark on the training sequence of the Middlebury [2] dataset with average end-point errors (px) in (a). In (b) we show an example of the low-quality SURF matches used to demonstrate the robustness to outliers. The matches are represented by their equivalent motion vector.



	1	2	3	4	5	6	Average
1	-	96.4 / 90.9 / 90.8	85.2 / 86.3 / 88.7	86.6 / 84.8 / 86.9	85.1 / 85.0 / 87.5	00.0 / 41.7 / 90.1	70.7 / 77.7 / 88.8
2	93.3 / 93.3 / 90.8	-	93.9 / 92.6 / 94.3	78.2 / 91.4 / 92.4	88.1 / 86.7 / 91.4	00.0 / 76.0 / 92.9	70.7 / 88.0 / 92.4
3	83.6 / 83.2 / 86.9	90.8 / 88.9 / 90.3	-	91.4 / 91.7 / 93.2	89.3 / 88.0 / 93.9	00.0 / 91.0 / 96.4	71.0 / 86.6 / 92.1
4	80.9 / 73.2 / 85.5	88.4 / 84.8 / 87.1	92.3 / 92.2 / 93.0	-	90.2 / 90.8 / 95.4	93.3 / 89.9 / 96.8	89.0 / 86.2 / 91.6
5	00.0 / 73.6 / 83.6	76.1 / 74.3 / 86.3	86.9 / 85.4 / 91.7	89.9 / 89.9 / 93.3	-	95.5 / 91.2 / 97.7	69.7 / 82.9 / 90.5
6	00.0 / 00.4 / 83.9	00.0 / 68.0 / 89.2	00.0 / 87.0 / 94.5	92.0 / 90.8 / 95.6	94.4 / 92.5 / 96.7	-	37.3 / 67.7 / 92.0

Figure 3. Evaluation on the herzjesu DAISY dataset. Diagonal depth maps are ground truth. The other depth maps are computed from the row/column images pairs, where the column images are the references. We used our method with segment matches. Estimated occlusions are colored in green. The table shows the percentage of outliers – error greater than 5% of the depth range – for our algorithm without matches / our algorithm with segment matches / Graph Cut with DAISY (results from [27]).

LDOF cannot recover too large motions, and that the feature-based method looses accuracy near the boundaries where there are no features. Our method allows us to take advantage of all features, without filtering, to obtain an accurate warp. Our results are even better with the unfiltered matches than with the filtered matches, which reveals a weakness of a separate step of outlier removal: it is difficult to find the balance between removing too many good matches or leaving mismatches.

Conclusion

In this work we introduced a general framework allowing us to greatly extends the scope of applicability of variational optical flow techniques. We combined a modern powerful discontinuity preserving regularizer with a robust direct data term and features integration. We generalized and extended the model of [9] to support any point features and introduced the novel concept of *a priori matches* to enable the use of

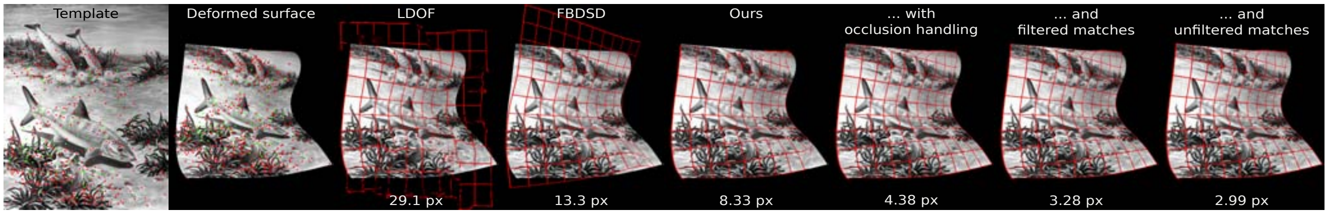


Figure 4. Results of deformable surface detection on synthetic data. The 613 unfiltered matches and the 103 filtered matches (green) are overlayed onto the template and target image. The images show the grid representing the inverse warp computed by each method as well as the average error of the warp. Best viewed on screen.

weakly localized features such as segments. Self-occlusion detection, rarely accounted for in optical flow estimation, further increases the robustness of our approach. This allowed us to showcase state of the art results on standard narrow-baseline optical flow and wide-baseline stereo. Preliminary results on non-rigid surface detection compare favorably with other methods and suggests promising use cases. Future work involves the improvement of each building block: higher-order regularization, richer direct data term and new features such as contours.

Acknowledgements. This research has received funding from the EU's FP7 through the ERC research grant 307483 FLEXABLE.

References

[1] Special session on robust optical flow, 2013. German Conference on Pattern Recognition.

[2] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 2011.

[3] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. *ECCV*, 2006.

[4] A. C. Berg and J. Malik. Geometric blur for template matching. In *CVPR*, 2001.

[5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 2001.

[6] G. Bradski. The OpenCV library. *Dr. Dobb's Journal of Software Tools*, 2000.

[7] K. Bredies. Recovering piecewise smooth multichannel images by minimization of convex functionals with total generalized variation penalty. *SFB Report*, 6, 2012.

[8] K. Bredies, K. Kunisch, and T. Pock. Total generalized variation. *SIAM Journal on Imaging Sciences*, 2010.

[9] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *PAMI*, 2011.

[10] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 2011.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[12] R. Garg, A. Roussos, and L. Agapito. A variational approach to video registration with subspace constraints. *IJCV*, 2013.

[13] V. Gay-Bellile, A. Bartoli, and P. Sayd. Direct estimation of nonrigid registrations with image-based self-occlusion reasoning. *PAMI*, 2010.

[14] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 2012.

[15] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, 1988.

[16] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 1981.

[17] C. Liu, J. Yuen, and A. Torralba. SIFT flow: Dense correspondence across scenes and its applications. *PAMI*, 2011.

[18] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[19] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *DARPA Image Understanding Workshop*, 1981.

[20] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, and X. Zhang. On building an accurate stereo matching system on graphics hardware. In *Third Workshop on GPUs for Computer Vision*, 2011.

[21] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Real time localization and 3D reconstruction. In *CVPR*, 2006.

[22] M. Perriollat and A. Bartoli. A computational model of bounded developable surfaces with application to image-based three-dimensional reconstruction. *Computer Animation and Virtual Worlds*, 2012.

[23] J. Pilet, V. Lepetit, and P. Fua. Fast non-rigid surface detection, registration and realistic augmentation. *IJCV*, 2008.

[24] D. Pizarro and A. Bartoli. Feature-based deformable surface detection with self-occlusion reasoning. *IJCV*, 2012.

[25] R. Ranftl, S. Gehrig, T. Pock, and H. Bischof. Pushing the limits of stereo using variational stereo estimation. In *Intelligent Vehicles Symposium (IV)*, 2012.

[26] E. Rosten, R. Porter, and T. Drummond. Faster and better: A machine learning approach to corner detection. *PAMI*, 2010.

[27] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *PAMI*, 2010.

[28] L. Wang, U. Neumann, and S. You. Wide-baseline image matching using line signatures. In *ICCV*, 2009.

[29] Z. Wang, F. Wu, and Z. Hu. MSLD: A robust descriptor for line matching. *Pattern Recognition*, 2009.

[30] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic Huber-L1 optical flow. In *BMVC*, 2009.

[31] J. Wills, S. Agarwal, and S. Belongie. A feature-based approach for dense segmentation and estimation of large disparity motion. *IJCV*, 2006.

[32] L. Xu, J. Jia, and Y. Matsushita. Motion detail preserving optical flow estimation. In *CVPR*, 2010.

[33] K. Yamaguchi, D. McAllester, and R. Urtasun. Robust monocular epipolar flow estimation. In *CVPR*, 2013.

[34] G. Yu and J.-M. Morel. ASIFT: An Algorithm for Fully Affine Invariant Comparison. *Image Processing On Line*, 2011.

[35] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *ECCV*, 1994.

[36] H. Zimmer, A. Bruhn, and J. Weickert. Optic flow in harmony. *IJCV*, 2011.