HAL
open science

A Variational Method for Scene Flow Estimation from Stereo Sequences

Frédéric Huguet, Frédéric Devernay

To cite this version:

Frédéric Huguet, Frédéric Devernay. A Variational Method for Scene Flow Estimation from Stereo Sequences. International Conference on Computer Vision, IEEE, Oct 2007, Rio de Janeiro, Brazil. 10.1109/ICCV.2007.4409000 . inria-00262139

HAL Id: inria-00262139

https://inria.hal.science/inria-00262139

Submitted on 11 Mar 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Variational Method for Scene Flow Estimation from Stereo Sequences

Frédéric Huguet Frédéric Devernay
PRIMA Team, INRIA Rhone-Alpes
Montbonnot, France
{frederic.huguet,frederic.devernay}@inrialpes.fr

Abstract

This paper presents a method for scene flow estimation from a calibrated stereo image sequence. The scene flow contains the 3-D displacement field of scene points, so that the 2-D optical flow can be seen as a projection of the scene flow onto the images. We propose to recover the scene flow by coupling the optical flow estimation in both cameras with dense stereo matching between the images, thus reducing the number of unknowns per image point. The use of a variational framework allows us to properly handle discontinuities in the observed surfaces and in the 3-D displacement field. Moreover our approach handles occlusions both for the optical flow and the stereo. We obtain a partial differential equations system coupling both the optical flow and the stereo, which is numerically solved using an original multi-resolution algorithm. Whereas previous variational methods were estimating the 3-D reconstruction at time t and the scene flow separately, our method jointly estimates both in a single optimization. We present numerical results on synthetic data with ground truth information, and we also compare the accuracy of the scene flow projected in one camera with a state-of-the-art single-camera optical flow computation method. Results are also presented on a real stereo sequence with large motion and stereo discontinuities. Source code and sample data are available for the evaluation of the algorithm.

1. Introduction

Scene flow was introduced by Vedula *et al*. [19, 20] as the 3-D vector field defined on the surfaces of a scene, describing the motion of each 3-D point between two time steps. It can be seen as an extension of optical flow to 3-D, but optical flow can also be seen as the projection of the 3-D scene flow onto the images, resulting in a 2-D vector field. Several methods propose to reconstruct scene flow from the observed optical flow in one or several cameras [20, 21], but the reconstruction step is either under- or over-constrained, and the different cameras may give non-consistent optical flows. To overcome these problems, we use a minimal parametrization of scene flow from the optical flow and the disparity of a stereo image sequence (this view-dependent description of scene flow is sometimes called *disparity flow* [9]). Since this parametrization is done in image space, the problem becomes close to an optical flow estimation problem with more unknown and more measures per image point.

A lot of research has been carried out on using variational methods to compute optical flow since the pioneer work by Horn and Schunck [3]. Some methods changed the regularization term in order to cope with the presence of discontinuities in the optical flow [7]. Recent work focused on reducing the computational cost of these variational methods, leading to real-time performance [6] or parallel implementation [5]. However, the best results in terms of accuracy were obtained by Brox *et al*. [4]: they avoid linearization of the different energy terms in the variational formulation by warping the image at time $t+1$ onto the image at time t, and the global energy is only linearized inside the minimization algorithm. That way, they get rid of the inaccuracies due to the approximation of the energy terms, especially the data term which had always been linearized since Horn and Schunck. This method is also robust to illumination changes, and somewhat handles discontinuities as well as occlusions, although the latter are not treated explicitly. A work by Slesareva *et al*. [17] adapted directly this variational formulation to estimate dense disparity maps.

Concerning the estimation of *scene flow* in a variational framework, one method that does both reconstruction and scene flow estimation was proposed by Pons *et al*. [14]. Scene flow estimation is performed by alternatively optimizing the reconstruction and the 3-D motion field. The latter is done by optimizing an energy that takes into account the difference between consecutive images re-projected on the computed 3-D reconstruction. Some recent works propose joint estimation of disparity and motion : the method by Dongbo Min *et al*. [13], which nevertheless misses illumination variations and occlusions handling, and the work by Isard and MacCormick [11] which only computes inte-

978-1-4244-1631-8/07/$25.00 ©2007 IEEE

ger disparity and flow values.

We propose a method that computes scene flow by joint estimation of the reconstructed surface and the motion field from a calibrated stereoscopic image sequence. This method takes into account the epipolar constraint between images taken at the same time, leading to a minimal parametrization of the scene flow. Only 4 variables are optimized at each pixel in the reference image: the stereo disparity at time t, the optical flow, and the disparity at time $t+1$ (the 3-D scene flow can be directly computed from these variables). This leads to a set of highly coupled non-linear partial differential equations which are solved by a multi-resolution algorithm. Our method avoids the linearization of the energy minimized by our algorithm. Indeed, it was numerically proved by Brox *et al*. that better results can be obtained by avoiding the linearization of the optical flow constraint. This is generalized to every constraint in our method. Besides the regularization terms can handle discontinuities both in the reconstruction and in the motion field, thus allowing fractures to appear on a smooth surface during time.

The rest of this paper is organized as follows: We first explain the mathematical formulation which couples optical flow and stereo, and the different terms of the energy that has to be minimized. We then expose the numerical difficulties tied to this problem, and the global algorithm. Finally, we present numerical results obtained on synthetic and real stereo sequences with the associated ground truth, and a real stereo sequence associated with a non rigid scene , with large motion and stereo discontinuities.

2. A unified variational formulation for optical flow and stereo

Our goal is to estimate a dense scene flow, while preserving the surfaces and motion discontinuities. Zhang and Kambhamettu [22] achieve this by first segmenting the scene, and then applying piecewise regularization, but that problem can also be solved by using an appropriate regularization functional.

Since we are working on a stereo image sequence, we first rectify the two image streams so that the stereo disparity is along the horizontal direction in the images. Gaussian smoothing ($\sigma = 1.25$) is also applied to the images in order to avoid numerical instabilities [2]. Our method uses the numerical benefits of the work by Brox *et al*.: robustness to changes in illumination thanks to the constant image gradient constraints, and robustness to stereo or optical flow occlusions by using the Ψ regularization function.

Let $I_l(x,y,t), I_r(x,y,t) : \Omega \subset \mathbb{R}^3$ be the left and right image sequences (Ω is the rectangular definition domain of the images). Let $(u,v) : \Omega \to \mathbb{R}^2$ be the optical flow in the left image, and $(d,d') : \Omega \to \mathbb{R}^2$ be the disparity maps at time t and at time $t+1$. $\mathbf{w} = (u,v,1)^\top$ is the displacement vector between the left image at time t and I_l at time $t+1$, $\mathbf{d} = (d,0,0)$ is the displacement between I_l and I_r at time t, and $\mathbf{d}' = (d',0,0)$ is the displacement between I_l and I_r at time $t+1$. As shown in Fig. 1, a point (x,y,t) in I_l corresponds to the points $(x+u(x,y), y+v(x,y), t+1)$ in I_l, $(x+d(x,y), y, t)$ in I_r, and $(x+u(x,y)+d'(x,y), y+v(x,y), t+1)$ in I_r: *the reference for the scalar functions u, v, d and d' is always I_l at time t*. It is clear that the 3-D reconstruction of the scene point observed at position (x,y) and time t in I_l can be obtained from d, and similarly its reconstruction at time $t+1$ is obtained from u, v, and d'. Scene flow can then easily be computed as the difference between these two positions.

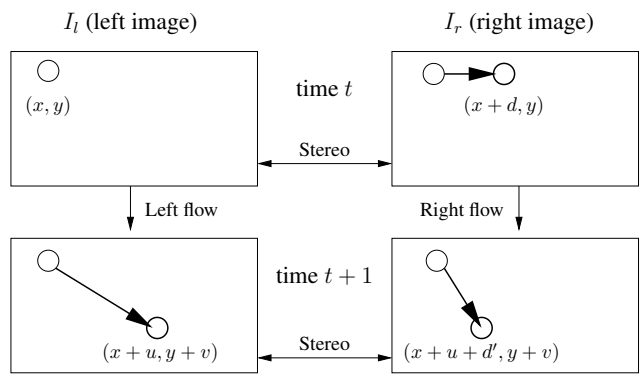

Figure 1. The motion of a projected scene point between two time steps as seen in the stereo images, and the associated functions.

We write the global energy as a sum of a data term and a regularization term:

$$E(u,v,d,d') = E_{Data} + \alpha E_{Smooth}, \tag{1}$$

α being the regularization parameter. E_{Data} is composed of four terms, corresponding to the four relations between images shown on Fig. 1:

$$E_{Data} = \int_\Omega \left(\beta_{fl}E_{fl} + \beta_{fr}E_{fr} + \beta_{st}E_{st} + \beta_s E_s\right) \mathbf{dx}. \tag{2}$$

$(x,y) : \Omega \to \beta_{fl}(x,y)$ is 1 for non occluded pixels for the left optical flow and 0 otherwise. The other β functions play a similar role for the occlusions associated with each part of E_{data}. Let us introduce the following notation for the difference in intensity and illumination between two image points:

$$\Delta(I,\mathbf{x};I',\mathbf{y}) = |I'(\mathbf{y}) - I(\mathbf{x})|^2 + \gamma|\nabla I'(\mathbf{y}) - \nabla I(\mathbf{x})|^2, \tag{3}$$

where $\nabla = (\partial_x, \partial_y)^\top$. The four terms in E_{Data} can be written:

$$E_{fl}(u,v,d,d') = \Psi\left(\Delta(I_l,\mathbf{x};I_l,\mathbf{x}+\mathbf{w})\right), \tag{4}$$

$$E_{fr}(u,v,d,d') = \Psi\left(\Delta(I_r,\mathbf{x}+\mathbf{d};I_r,\mathbf{x}+\mathbf{w}+\mathbf{d}')\right), \tag{5}$$

$$E_{st}(u,v,d,d') = \Psi\left(\Delta(I_l,\mathbf{x}+\mathbf{w};I_r,\mathbf{x}+\mathbf{w}+\mathbf{d}')\right), \tag{6}$$

$$E_s(u,v,d,d') = \Psi\left(\Delta(I_l,\mathbf{x};I_r,\mathbf{x}+\mathbf{d})\right). \tag{7}$$

E_{fl} is the data term corresponding to the left optical flow, and E_{fr} corresponds to the right optical flow, which has the same vertical component as the left optical flow. Similarly, E_s corresponds to stereo matching between the left and right images at time t, and E_{st} corresponds to stereo matching at time $t+1$. Pixels in the left image may become occluded in some of the other three images, and quadratic penalizers would give them too much influence on the solution. To solve this problem, we use the Ψ function [1, 4], defined by $\Psi(s^2) = \sqrt{s^2+\epsilon^2}$ (with $\epsilon = 0.001$), which leads to a robust energy, corresponding to L^1 minimization, but is still differentiable everywhere. The Ψ function is applied separately to each data term, since pixels may be occluded by stereo, but not by optical flow, and vice-versa. Besides, eq. (3) incorporates a gradient constancy assumption in all data terms [4], so that the energy is also robust to illumination changes (local or global) and non-Lambertian surfaces (the stereo terms may be highly affected by such surfaces, since they use images coming from different viewpoints). The γ parameter should be set empirically, depending on how much illumination change is expected in the scene.

We could have considered that the disparity d at time t is given from the previous scene flow estimation (between time $t-1$ and time t), but if the estimated disparity d contains errors, theses errors would propagate to d', u, and v. By minimizing the four data terms, we will be able to re-evaluate all the components of the scene flow: The 3-D reconstruction (from d), and the 3-D motion field (from u, v, and $d'-d$).

The regularization term is:

$$E_{Smooth} = \int_\Omega \Psi(|\nabla u|^2 + |\nabla v|^2 + \lambda|\nabla(d'-d)|^2 + \mu|\nabla d|^2)\mathbf{dx}. \quad (8)$$

By reducing the influence of high gradients of the optical flow or the disparity on the global energy, the Ψ function has a different role here: it helps preserving the discontinuities of the functions u, v, d, and d' [6]. Unlike in the data term, Ψ is applied to the sum of the gradient norms, since discontinuities usually appear simultaneously in the disparity d, the optical flow (u, v), and the disparity flow $d'-d$ (except in some special cases, as in the synthetic example used in the results below).

The effect of the regularization on the 3-D scene flow should not depend on the orientation of the motion field with respect to the camera, so the λ parameter should be set properly to scale optical flow versus disparity flow, but should not be greater than μ to avoid oscillations during optimization: $\lambda < h/b$, where h is the average distance from the cameras to the scene and b is the baseline of the stereo setup. The effect of this parameter will be more regular disparity flow $(d'-d)$ and smaller discontinuities when the baseline is smaller. The μ parameter tunes the relative weight between the initial disparity and the optical flow. Since the typical discontinuities in both terms observed on the scene should have the same effect on E_{Smooth}, a good guess is $\mu = hs/bS$ where s is the typical expected magnitude (in world units) of the 3-D scene flow, and S a typical size of the scene: if the typical motion between t and $t+1$ is small with respect to the size of the scene, then μ should be small too.

3. Optimization

3.1. Euler-Lagrange equations

According to calculus of variations, an extremum of the total energy E satisfies the four Euler-Lagrange equations, $\nabla E(u, v, d, d') = 0$, which can be rewritten as $(\partial_u E, \partial_v E, \partial_d E, \partial_{d'} E) = (0, 0, 0, 0)$. While being necessary, this condition is not sufficient, and the solutions to the Euler-Lagrange equations may also be local extrema of eq. (1). We will see later how a multi-resolution approach helps solving this problem.

The four equations can be computed the same way, using the variational calculus tools, and have similar terms. Let us introduce the following abbreviations:

$$I_{lx} := \partial_x I_l(\mathbf{x}+\mathbf{w}), \quad I_{lxz} := \partial_x I_l(\mathbf{x}+\mathbf{w}) - \partial_x I_l(\mathbf{x}), \quad (9)$$

$$I_{ly} := \partial_y I_l(\mathbf{x}+\mathbf{w}), \quad I_{lyz} := \partial_y I_l(\mathbf{x}+\mathbf{w}) - \partial_y I_l(\mathbf{x}), \quad (10)$$

$$I_{lz} := I_l(\mathbf{x}+\mathbf{w}) - I_l(\mathbf{x}), \quad I_{lyy} := \partial^2_{yy} I_l(\mathbf{x}+\mathbf{w}), \quad (11)$$

$$I_{lxx} := \partial^2_{xx} I_l(\mathbf{x}+\mathbf{w}), \quad I_{lxy} := \partial^2_{xy} I_l(\mathbf{x}+\mathbf{w}), \quad (12)$$

$$I_l^{t+1} := I_l(\mathbf{x}+\mathbf{w}) \quad (13)$$

and similar abbreviations for the right image I_r, as well as the following scalars: The last of the previous notation is useful to see the time index in the following equations.

$$\Psi'_{fl} = \partial_x \Psi(\Delta(I_l, \mathbf{x}; I_l, \mathbf{x}+\mathbf{w})) \quad (14)$$

$$\Psi'_{fr} = \partial_x \Psi(\Delta(I_r, \mathbf{x}+\mathbf{d}; I_r, \mathbf{x}+\mathbf{w}+\mathbf{d}')) \quad (15)$$

$$\Psi'_{st} = \partial_x \Psi(\Delta(I_l, \mathbf{x}+\mathbf{w}; I_r, \mathbf{x}+\mathbf{w}+\mathbf{d}')) \quad (16)$$

$$\Psi'_{\text{div}} = \partial_x \Psi(|\nabla u|^2 + |\nabla v|^2 + \lambda|\nabla(d'-d)|^2 + \mu|\nabla d|^2). \quad (17)$$

By computing $\partial_u E$ using a Gâteaux derivative, we obtain:

$$\begin{aligned}
&\beta_{fl}\Psi'_{fl}\cdot\big(I_{lx}I_{lz} + \gamma(I_{lxx}I_{lxz} + I_{lxy}I_{lyz})\big) + \\
&\beta_{fr}\Psi'_{fr}\cdot\big(I_{rx}I_{rz} + \gamma(I_{rxx}I_{rxz} + I_{rxy}I_{ryz})\big) + \\
&\beta_{st}\Psi'_{st}\cdot\big((I_r^{t+1} - I_l^{t+1})(I_{rx} - I_{lx}) + \gamma\big((I_{rx} - I_{lx})(I_{rxx} - I_{lxx}) + \\
&\quad (I_{ry} - I_{ly})(I_{rxy} - I_{lxy})\big)\big) - \alpha\,\text{div}(\Psi'_{\text{div}}\nabla u) = 0, \quad (18)
\end{aligned}$$

This equation is composed of a data term, coming from E_{Data}, and a diffusion term in which occurs the divergence operator. By lack of space we will not show the three other equations coming from $\partial_v E = 0$, $\partial_{d'} E = 0$, $\partial_d E = 0$ but are similar to the latter.

The boundary conditions for our problem are the Neumann conditions: $\forall f \in \{u, v, d, d'-d\}, \nabla f \cdot \mathbf{n} = 0$, where $\mathbf{n}$ is the external normal to the borders of image I_l.

In this system of partial differential equations, the four unknown functions of our system, u, v, d and d', are highly coupled, but solving these equations will lead to scene flow reconstruction.

3.2. Numerical solution

As we explained, the energy is not trivially convex, since the optical flow constraint was not linearized, and the non-linearities are present both in the data term and in the diffusion term of the Euler Lagrange equations. This makes the problem ill-posed, and we cannot use gradient descent to minimize the energy as in [15]. In order to solve these highly non-linear coupled differential equations, we use an incremental multi-resolution algorithm, with fixed-point iterations on the solution(u, v, d, d') to improve it at each resolution level. A similar method was proposed by Brox *et al.* [4] to solve the optical flow problem. The stereo image pyramids are computed with a down sampling factor η, $0.5 < \eta < 1$ to get a smooth transition between pyramid levels (we used $\eta = 0.9$). The multi-resolution approach ensures that we converge to a global minimum, as demonstrated in [12]. This algorithm has been shown to work on many problems, and was recently improved to get near real-time performance [6].

The data term in the Euler Lagrange equations, *e.g.* in eq. (18), is made of image values and image gradients computed with respect to the reference image I_l at time t. This is equivalent to warping the three other images (I_l at time $t+1$, and I_r at times t and $t+1$) from the same pyramid level onto image I_l at time t, using the current solution (u, v, d, d'), and computing the data terms from these warped images and their gradients.

We deal with the non-linearities of the equations at a given pyramid level by using two nested fixed point iterations, obtained by doing a first order Taylor expansion of the Euler Lagrange equations to transform it into a linear system. The inside fixed point iterations compute small increments of the solution (du, dv, dd, dd'), and the images are re-warped using $(u + du, v + dv, d + dd, d' + dd')$ at each iteration. The outside fixed point iterations update the full solution (u, v, d, d'). We refer to sec. 3.2 of [4] for full details on how to compute the fixed point iterations from the Euler Lagrange (although the referred article concerns the simpler optical flow problem). The inside fixed point iterations uses the SOR method to solve the final linear system (described in [10]). In this method, the system matrix is separated in three parts: the diagonal, and the upper and lower triangular sub-matrices. Consequently, different orderings of the lines and columns of the system will yield different results at each iteration. Our implementation uses alternatively four different orderings, where the image pixels are scanned in four different directions, in order to reduce the asymmetry induced by each individual SOR iteration, which is not visible on the optical flow problem, but induced oriented waves in the scene flow numerical solution.

The stopping conditions for the two fixed point iterations are measured from the relative L^2 norm between consecutive increments. We used 0.05 as the stopping condition for the inner fixed point iterations, and 0.01 for the outer fixed point iterations.

Once the optimization is obtained at a given pyramid level, that solution is scaled by $1/\eta$, up-sampled to the next resolution level, and the same process is repeated until the full resolution is reached.

3.3. Occlusions computation

Occlusions are handled by computing the functions β_{fl}, β_{fr}, β_{st}, β_s at each beginning of the outer fixed point iteration. So we take into account each $(u, v, d, d^{'})$ increment to compute the occlusions maps. The β functions take the 1 value for pixels non occluded and 0 otherwise, so that for pixels occluded everywhere, only the regularization term is kept.
We describe the steps of β_s estimation, the other functions are computed using a similar principle:

- We first warp the disparity d to the right image $I_r(., t)$ using Z buffering.
- We backwarp this disparity map to the left image $I_l(., t)$, and we add a tolerance (1.5 pixel) to the remapped disparity.
- We compute the occlusion map by comparing d with the backwarped disparity with its tolerance.

The update of E_{data} for each pixel of the reference image is then realized.

3.3.1 Full algorithm with initialization

Since the problem to solve is strongly non-linear and non-convex, it must be carefully initialized in order to avoid local minima which correspond to a wrong solution.

For the optical flow problem [4], and especially when using a multi-resolution algorithm, the coarsest resolution can be as small as possible, and the optical flow is usually initialized to 0. The reason for this choice is that the optical flow is usually small compared to the image dimensions, and it is easy to find a reasonable image resolution such that the scaled down optical flow is below 0.5 pixel, which is usually enough to ensure convergence to the global minimum.

In the scene flow case, we have a mixed problem: it looks like optical flow if we consider each camera separately, but we simultaneously try to solve a stereo problem between the left and right images. The characteristics of the stereo

problem are very different from those of the optical flow: the amplitude of the stereo disparity is usually comparable to image size (it is usually even bigger than the size of the objects as seen in the images), and there are lots of occluded areas. For these reasons, many multi-resolution approaches usually fail on stereo if they start at a very coarse resolution, and our method will probably equally fail in that situation.

Consequently, we chose to start the scene flow algorithm at an intermediate resolution, and to initialize the four functions (u, v, d, d') with non-zero values. First, we initialize d using a stereo algorithm [8] which computes the disparity from the highest resolution images (level 1 of the pyramid). The disparity error of a given stereo algorithm can be easily evaluated using standard benchmarks [16], and we compute the pyramid level b such that the downscaled nominal disparity error is below 0.5 pixels. We also compute a pyramid level a, which is higher (*i.e.* coarser) than b, so that the expected optical flow at this level is below 0.5 pixels. We then solve the optical flow problem – by keeping the terms of eq. (1) dealing with the left optical flow, which bring us back to [4] – for the left and the right images separately, from level a to level b, and we obtain estimates for the left optical flow (u, v) and the right optical flow (u', v'). The initial disparity d is then refined from a level c to the level b (c is often chosen as being equal to b, but can be chosen by the user), using the same method and keeping only the terms dealing with stereo at time t. d' is initialized by adding the difference between u' and u to d, and warping the result to I_l at time t (details are given in Algorithm 1).

Finally, the scene flow estimation algorithm is applied to the four images, from level b to level 1 of the pyramid. The full scene flow estimation algorithm, including the initialization phase, is detailed in Algorithm 1.

Algorithm 1 Full scene flow estimation

Ensure: Compute scene flow (u, v, d, d') from t and $t+1$ stereo pyramids (each pyramid has a levels)
Require: $a, b, c \in \mathbb{N}, a > b >= 1, a > c >= b >= 1$
  $u \leftarrow 0, v \leftarrow 0, u' \leftarrow 0, v' \leftarrow 0$
  for $l = a$ to b **do**
    $(u, v) \leftarrow$ left optical flow from (u, v) and level l
    $(u', v') \leftarrow$ right optical flow from (u', v') and level l
  end for
  $d \leftarrow$ stereo from [8]
  for $l = c$ to b **do**
    $d \leftarrow$ disparity at time t from d and level l
  end for
  $d^{'}(\mathbf{x} + (u, v)) \leftarrow d + u'(\mathbf{x} + d) - u$
  for $l = b$ to 1 **do**
    $(u, v, d, d') \leftarrow$ scene flow from (u, v, d, d') and level l
  end for

4. Results and evaluation

Whereas there are numerous datasets with ground truth for various algorithms in computer vision, the scene flow problem is probably not mature enough to deserve a proper evaluation benchmark. However, such datasets exist for sub-problems of the scene flow: optical flow and stereo.

The standard benchmark for optical flow is the Yosemite sequence, a flight sequence on a ray-tracer-rendered landscape, with flow and depth ground truth. Unfortunately, at that time a single camera was rendered for the sequence, and though a second camera could be rendered by using the depth to warp the first image, the quality would be low, and the occluded areas would be missing.

For the stereo problem, several datasets are available, each consisting in 8 view of the same scene, where all the optical centers are aligned and evenly spaced, and the images are rectified [16]. Incidentally, these images can be used to benchmark a scene flow algorithm: imagine a set of two rectified cameras which observe a static scene, and are translated along the straight line joining their optical centers. All those images are present in the stereo benchmark datasets. However, they represent a special configuration for the scene flow estimation, since the optical flow part is strictly horizontal ($v = 0$), and the disparity maps are the same ($d' = d$), but since our algorithm doesn't know anything about these, it is still a good benchmark. We took images 2 and 6 of the Venus, Teddy and Cones datasets as the stereo pair at time t, and images 4 and 8 as the stereo pair at time $t + 1$. Ground truth is given as the disparity from 2 to 6, and the optical flow is half the disparity.

In order to evaluate our algorithm on a more general scene flow, we also generated synthetic images of a rotating sphere (Fig. 2). This scene represents the extreme case where a 3-D reconstruction will not give any information about what is happening in the scene, and all the information is contained in the scene flow: since the sphere is rotating, the reconstruction remains identical over time. Besides, the hemispheres are rotating in opposite directions, which generates a strong discontinuity in the scene flow, and we will be able to check if the method properly recovers that discontinuity.

The evaluation is done by computing the RMS error on the four maps u, v, d, d'. The optical flow maps (u, v) are evaluated together, and the disparity maps are evaluated separately: although they are measured also in pixels, measuring the disparity is more difficult because of the disparity range and the occlusions. Results of these evaluations are shown Fig. 3. Fig. 4 compares the angular error of the optical flow components of scene flow, compared to optical flow computed using our method or [4]. Fig. 5 shows the resulting u, v, d and d' maps for the ball example, showing that the discontinuity was properly handled by our algorithm, and the generated occlusion maps. Figures 6 and 7 present results on a real stereo pair sequence with large motion and discontinuities.

For further evaluation of our method, we also provide an

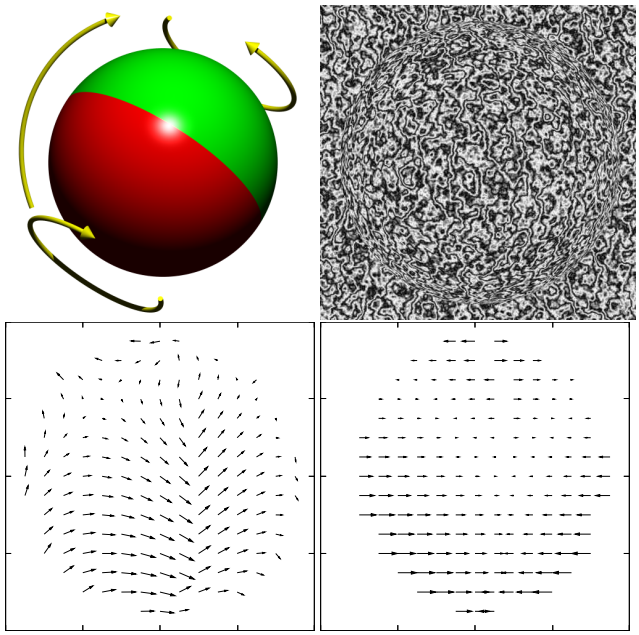

Figure 2. The synthetic sample scene is a rotating textured sphere, where the two hemispheres rotate separately (top-left, image on top-right)). The 3-D reconstruction remains unchanged by these rotations The 3-D motion information is only measurable from the scene flow: (u, v) (bottom left) and $d' - d$ (bottom right) show a scene flow discontinuity allow the vertical meridian.

Dataset	(u, v)	d	d'
Venus	0.31	0.97	1.48
Teddy	1.25	2.27	6.93
Cones	1.11	2.11	5.24
Sphere	0.69	3.73	3.81

Figure 3. RMS error in pixels on the four maps computed by our scene flow algorithm with the different datasets.

Dataset	μ_{of}	σ_{of}	μ_{sf}	σ_{sf}
Venus	1.06	1.17	0.98	0.91
Teddy	0.43	0.49	0.51	0.66
Cones	0.66	1.21	0.69	0.77
Sphere	1.50	5.65	1.75	6.07

Figure 4. Mean μ_{sf} and standard deviation σ_{sf} of the absolute angular error in degrees of the optical flow component (u, v) of the scene flow, compared to the angular error (μ_{of}, σ_{of}) of the optical flow computed separately.

OpenCV based implementation of the algorithm, with the sphere dataset[1] (other datasets can be downloaded from the Middlebury stereo page). The sample code can be used to compute optical flow, stereo, or scene flow, using the unified approach presented in this paper.

5. Conclusion

In this paper, we presented a variational framework to compute scene flow from a stereoscopic image sequence.

[1]The source code is included in the additional material, and is available on http://devernay.free.fr/vision/varsceneflow/

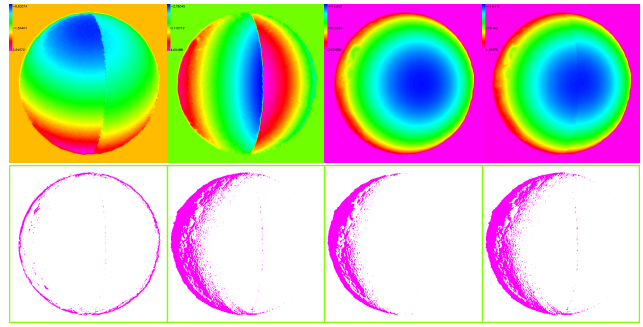

Figure 5. Top: the recovered u, v, d, d' maps for the ball example ($-7 < u < 4$, $-4 < v < 4$, $-113 < d < 1$, $-115 < d' < 2$). Notice the vertical discontinuity in d', due to the fact that the reference coordinates are those of the left image at time t. Bottom: the occlusion maps for the data terms corresponding to left flow, right flow, disparity at t and disparity at $t + 1$.

This method couples optical flow estimation with dense stereo matching by minimizing a global energy. The method handles discontinuities in the 3-D geometry or in the 3-D motion vector field, is robust to the illuminations changes and moreover handles the occlusions due to optical flow and stereo.

Our method extends the work made by Brox *et al.* [4] on accurate optical flow estimation, by adding constraints due to the epipolar geometry, and we showed that the same kind of numerical solution can be used to solve both problems. However, the nature of a disparity map is different from the optical flow, in the sense that occlusions are larger, and the disparity range is comparable to the size of the objects in the image, causing many difficulties to many multi-resolution stereo algorithms. We thus proposed a two-step algorithm, where the initial solution is bootstrapped by separate solutions to the optical flow and the stereo problem, and that initial solution is then refined by our scene flow estimation method. This is the first paper on scene flow which presents a quantitative evaluation of the method, by comparing the optic flow component of the scene flow with the results of the most accurate variationnal optical flow method to our knowledge. Moreover, our experiments showed that the method is able to handle real stereo sequences with large motion and stereo discontinuities.

In the near future, we expect to have a mathematical proof for the convergence of this method, and we will also work on speeding up the algorithm, probably by porting some recent work on near real-time variational methods [6] to solve the scene flow problem. Moreover, we would like to estimate a deterministic continuous function for the β coefficients handling discontinuities. Previous work uses a probabilistic formulation [18], but a deterministic continuous approach could be better integrated into our variational formulation.

Figure 6. An example with real data (images are 854×854 pixels). The time interval between the top and the bottom stereo pair is 1.5s, resulting in illumination variations, large motion (both in translation and rotation), and a clear motion discontinuity in the mouth region. The ranges in pixels for the scene flow components on this sample set are $u \in [-131, 1]$, $v \in [-49, 33]$, $d, d' \in [-122, -39]$.

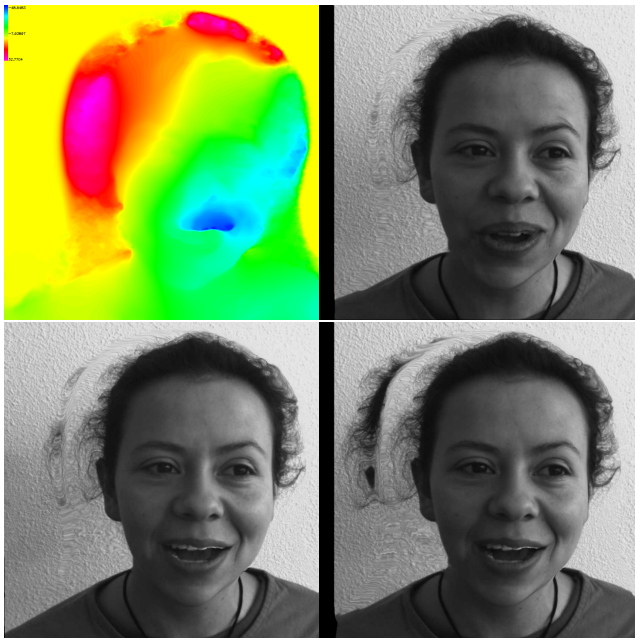

Figure 7. Sample results on real data: on top-left, the vertical flow component of scene flow shows clearly that the mouth discontinuity was recovered. The right image at time 0 (top-right) and the stereo pair at time 1.5s were warped to the left image at time 0, showing where the scene flow was correctly estimated. Although the discontinuity in the mouth area was recovered, notice that the scene flow is not pixel-accurate in this area.

References

[1] G. Aubert and P. Kornprobst. A mathematical study of the relaxed optical flow problem in the space BV(ω). *SIAM J. Math. Anal.*, 30(6):1282–1308, 1999.

[2] J. Barron, D. Fleet, S. Beauchemin, and T. Burkitt. Performance of optical flow techniques. In *Proc. IEEE CVPR*, pages 236–242, 1992.

[3] B.Horn and B.Schunck. Determining optical flow. *Artifical Intelligence*, 17:185–203, 1981.

[4] A. Brox, N. Bruhn, J. Papenberg, and T. Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proc. 8th ECCV*, volume 3024 of *LNCS*, pages 25–36, Prague, Czech Republic, 2004. Springer-Verlag.

[5] A. Bruhn, J. Weickert, C. Feddern, T. Kohlberger, and C. Schnörr. Variational optical flow computation in real time. *IEEE Trans. Image Processing*, 14(5):608–615, 2005.

[6] A. Bruhn, J. Weickert, T. Kohlberger, and C. Schnorr. Discontinuity preserving computation of variational optic flow in real-time. In *ScaleSpace05*, pages 279–290, 2005.

[7] R. Deriche, P. Kornprobst, and G. Aubert. Optical-flow estimation while preserving its discontinuities: A variational approach. In *Proc. ACCV*, pages 71–80, 1995.

[8] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *International Journal of Computer Vision*, 70(1), Oct. 2006.

[9] M. Gong and Y.-H. Yang. Disparity flow estimation using orthogonal reliability-based dynamic programming. In *Proc. 18th ICPR*, pages 70–73. IEEE, 2006.

[10] F. Huguet and F. Devernay. A variational method for scene flow estimation from stereo sequences. Research Report 6267, INRIA, Aug. 2007.

[11] M. Isard and J. MacCormick. Dense motion and disparity estimation via loopy belief propagation. In *ACCV06*, pages II:32–41, 2006.

[12] M. Lefébure and L. D. Cohen. Image registration, optical flow and local rigidity. *J. Math. Imaging Vis.*, 14(2):131–147, 2001.

[13] D. Min and K. Sohn. Edge-preserving simultaneous joint motion-disparity estimation. In *ICPR06*, pages II: 74–77, 2006.

[14] J.-P. Pons, R. Keriven, and O. Faugeras. Modelling dynamic scenes by registering multi-view image sequences. In *Proc. IEEE CVPR*, volume 2, pages 822–827, 2005.

[15] J.-P. Pons, R. Keriven, O. Faugeras, and G. Hermosillo. Variational stereovision and 3D scene flow estimation with statistical similarity measures. In *Proc. IEEE ICCV*, page 597, 2003.

[16] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002.

[17] N. Slesareva, A. Bruhn, and J. Weickert. Optic flow goes stereo: A variational method for estimating discontinuity-preserving dense disparity maps. In *DAGM05*, page 33, 2005.

[18] C. Strecha, R. Fransens, and L. J. V. Gool. A probabilistic approach to large displacement optical flow and occlusion detection. In *ECCV Workshop SMVP*, pages 71–82, 2004.

[19] S.Vedula, S.Baker, P.Rander, R.Collins, and T.Kanade. Three-dimensional scene flow. In *Proc. IEEE ICCV*, pages 722–729, 1999.

[20] S. Vedula and S. Baker. Three-dimensional scene flow. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(3):475–480, 2005.

[21] Y. Zhang and C. Kambhamettu. Integrated 3D scene flow and structure recovery from multiview image sequences. In *Proc. IEEE CVPR*, pages II: 674–681, 2000.

[22] Y. Zhang and C. Kambhamettu. On 3D scene flow and structure estimation. In *Proc. IEEE CVPR*, page 778, 2001.