Local/Global Scene Flow Estimation

Julian Quiroga, Frédéric Devernay, James L. Crowley

LOCAL/GLOBAL SCENE FLOW ESTIMATION

Julian Quiroga *Frédéric Devernay* *James Crowley*

PRIMA team, INRIA Grenoble Rhone-Alpes, France
{julian.quiroga,frederic.devernay,james.crowley}@inria.fr

ABSTRACT

The scene flow describes the 3D motion of every point in a scene between two time steps. We present a novel method to estimate a dense scene flow using intensity and depth data. It is well known that local methods are more robust under noise while global techniques yield dense motion estimation. We combine local and global constraints to solve for the scene flow in a variational framework. An adaptive TV (Total Variation) regularization is used to preserve motion discontinuities. Besides, we constrain the motion using a set of 3D correspondences to deal with large displacements. In the experimentation our approach outperforms previous scene flow from intensity and depth methods in terms of accuracy.

Index Terms— Scene flow, 3D motion, depth data, variational

1. INTRODUCTION

Capturing the 3D motion of a scene is a topic of great interest in computer vision. The scene flow is defined as the 3D motion field of the scene and its computation can be performed using data provided from different sources, e.g., color cameras and depth sensors. Motion in the form of scene flow provides powerful cues for visual systems. However, the scene flow of non-rigid scenes cannot be estimated from a single view without additional assumptions or information, and it requires a fully calibrated stereo or multi-view camera system, which is not always available. With the arrival of depth cameras it has been possible to compute scene flow using a registered sequence of depth and intensity images.

There are different choices when computing scene flow using intensity and depth: *i)* inferring the scene flow using depth after a 2D motion estimation, *ii)* computing the scene flow in 3D performing or not surface reconstruction or *iii)* computing the 3D motion using intensity and depth in the image domain. In the first case, the 2D motion field can be estimated using an accurate optical flow algorithm. Total variation (TV)-L^1 based method has proven to be the most effective [1] and simpler versions can run in real time [2]. Using the depth information, the scene flow can be generated by back-projecting 2D motions. However, optical flow computation has inherent challenges that can affect the 3D motion estimation. The regularization employed to fill the estimation on untextured regions can degrade the motion discontinuities of the flow. Besides, the optical flow is solved to be consistent with the observed brightness data which may not be enough to explain the 3D motion field. In this approach, depth data is not used to estimate the 2D motion. On the other hand, if computation is performed in 3D, intensity and depth data are used to generate a set of 3D points. Motion can be estimated by reconstructing a surface, e.g., by generating a triangular mesh, or by representing the surface directly as a point cloud. The later approach is computationally more efficient and has proved to perform correct 3D motion field estimation [3]. However, in this unstructured representation, the set of 3D motions hypotheses can be unnecessary large. Besides, the lack of a 3D structure makes the scene flow estimation more sensitive to noisy and missing data.

In this work, we estimate a dense scene flow using intensity and depth. We combine local and global constraints to solve for the scene flow in a variational framework. Inspired by [4] we locally constrain the scene flow in the image domain but instead of solving for a local motion we get a dense scene flow by performing an adaptive TV regularization. This way, we are able to estimate an accurate dense scene flow while preserving motion discontinuities. Besides, we include a set of 3D correspondences to deal with large displacements. Our formulation supports different motion models and local/global trade off can be adjusted to control the local-rigidity assumption.

2. RELATED WORK

Since the introduction of the scene flow [5], several approaches have been proposed to solve this problem. Most of them solve for 3D structure and 3D motion by using the data provided by a stereo or a multi-view camera system. The most intuitive way to compute scene flow is to reconstruct it from several optical flows [6]. However, it is difficult to recover a scene flow compatible with several observed optical flows which may be contradictory. Some authors introduce constraints of a full calibrated stereo structure [7, 8, 9, 10].

When depth data is provided, the 3D structure estimation is not needed anymore and both intensity and depth can be used for the motion estimation. Spies *et al.* [11] solve for optical and range flows: In that work depth data is used as an extra channel and the classical optical flow equation is adapted to constrain the observed depth data. Lukins and Fisher [12] extend this approach to multiple color channels and one aligned depth image. In both approaches the 3D motion field is computed by constraining the flow in intensity and depth images of an orthographically captured surface, so that the range flow is not used to support the 2D motion estimation. In contrast, in our work depth data is used in two ways: to model the image motion and to constrain the scene flow. A very close work is [13], which uses a locally rigid assumption to compute a local scene flow. However, this local approach can not be applied in untextured regions and there is no criterium for selecting good regions to be considered. Instead, we solve for a dense scene flow by regularizing the 3D motion field. Hadfield and Bowden [3] estimate scene flow by modeling moving points in 3D using particle filtering. This method is computationally expensive since a large set of motion hypothesis must be tested for each 3D point. We exploit the 2D parametrization of the data to formulate an efficient motion exploration where best optical flow practices can be used. Previous methods require a lot of computation time which limits their use in real-time applications. Instead, we introduce an auxiliary variable which allows a more efficient solution by alternating between iterative reweighted least squares (IRLS) and a TV solver based on the dual-ROF model [14].

3. LOCALLY RIGID MOTION

Let $\mathbf{X} = (X, Y, Z)$ be a 3D point in the camera reference frame at time $t-1$ and $\mathbf{X}' = (X', Y', Z')$ its location at time t. Let $\mathbf{x} = (x, y) = (X/Z, Y/Z)$ be the projection of $\mathbf{X}$ on the image, where for brevity we suppose unit focal lengths and optical center at the image origin. If $\mathbf{x}' = (x', y')$ is the projection of $\mathbf{X}'$, the *image flow* (u, v) induced by 3D motion $\mathbf{v} = \{v_X, v_Y, v_Z\}$ is given by:

$$u = x' - x = \left(\frac{X + v_X}{Z + v_Z} - \frac{X}{Z}\right) = \frac{1}{Z}\left(\frac{v_X - x v_Z}{1 + v_Z/Z}\right) \quad (1)$$

and (2)

$$v = y' - y = \left(\frac{Y + v_Y}{Z + v_Z} - \frac{Y}{Z}\right) = \frac{1}{Z}\left(\frac{v_Y - y v_Z}{1 + v_Z/Z}\right).$$

Using a Taylor series in the denominator term containing v_Z, we get

$$\left(\frac{1}{1 + v_Z/Z}\right) = \left(1 - \frac{v_Z}{Z} + \left(\frac{v_Z}{Z}\right)^2 - ...\right). \quad (3)$$

If $|v_Z/Z| \ll 1$, only the zero-order term remains and the image flow induced by $\mathbf{t}$ on a pixel $\mathbf{x}$ is given by:

$$\begin{pmatrix} u(\mathbf{x}, \mathbf{v}) \\ v(\mathbf{x}, \mathbf{v}) \end{pmatrix} = \frac{1}{Z}\begin{pmatrix} 1 & 0 & -x \\ 0 & 1 & -y \end{pmatrix}\begin{pmatrix} v_X \\ v_Y \\ v_Z \end{pmatrix}. \quad (4)$$

We also assume that the scene is composed of rigidly, but independently, moving 3D parts. Let $\mathbf{x_0}$ be the projection of 3D point $\mathbf{X_0}$ which has an associated 3D motion vector $\mathbf{v}$. The locally rigid assumption claims a 2D neighbourhood $N(\mathbf{x_0})$ where pixels move following the 3D motion $\mathbf{v}$. We define the warp function

$$\mathbf{W}(\mathbf{x}; \mathbf{v}) = \mathbf{x} + \begin{pmatrix} u(\mathbf{x}, \mathbf{v}) \\ v(\mathbf{x}, \mathbf{v}) \end{pmatrix} \quad (5)$$

which maps each $\mathbf{x} \in N(\mathbf{x_0})$ to its new position after 3D motion $\mathbf{v}$.

4. SCENE FLOW MODEL

Using simultaneously intensity and depth we state the scene flow computation as minimization problem of the energy

$$E(\mathbf{v}) = E_D(\mathbf{v}) + \alpha E_M(\mathbf{v}) + \beta E_R(\mathbf{v}), \quad (6)$$

where $\mathbf{v} = \{v_X, v_Y, v_Z\}$ denotes the motion field to be estimated. The first term $E_D(\mathbf{v})$ is the *data term* which measures how consistent is the estimated scene flow with the observed intensity and depth. To deal with large motion we include a *sparse matching term* $E_M(\mathbf{v})$ enforcing the flow to agree with the 3D motion of a set of interest points. Finally, the *regularization term* $E_R(\mathbf{v})$ is based on an adaptive TV on each component of the 3D motion field, favoring locally rigid motion and preserving motion discontinuities.

4.1. Data term

In our formulation we look for the scene flow $\mathbf{v}$ that best explains the intensity and depth data. We use the *brightness constancy assumption* (*BCA*) given by $I_2(\mathbf{W}(\mathbf{x}; \mathbf{v})) = I_1(\mathbf{x})$, where the warp function $\mathbf{W}(\mathbf{x}; \mathbf{v})$ maps each pixel $\mathbf{x}$ from I_1 to I_2 accordingly to the scene flow $\mathbf{v}$, see Eq. (5). Using this single equation only 2D motions can be estimated and even these are not uniquely computed (aperture problem). In fact, it is only possible to compute the component parallel to the image gradient wherever it is not vanishing. Since we are provided with a registered depth image, we include a *depth velocity constraint* (*DVC*) given by $Z_2(\mathbf{W}(\mathbf{x}; \mathbf{v})) = Z_1(\mathbf{x}) + v_Z(\mathbf{x})$, where $v_Z(\mathbf{x})$ is the Z component of the 3D motion of pixel $\mathbf{x}$. This equation enforces the consistency between the motion captured by the depth sensor and the estimated motion.

In order to cope with outliers brought by noise, occlusions or motion inconsistencies, a robust norm is required. We use the Charbonnier penalty $\Psi\left(s^2\right) = \sqrt{s^2 + \varepsilon^2}$ which is a differentiable approximation of the L^1 norm. This penalizer is applied separately to BCA and DVC, and the data term can be written as

$$E_D(\mathbf{v}) = \sum_{\mathbf{x}} \Psi\left(|\rho_I(\mathbf{x}, \mathbf{v})|^2\right) + \lambda \Psi\left(|\rho_Z(\mathbf{x}, \mathbf{v})|^2\right), \quad (7)$$

where ρ_I and ρ_Z are residuals given by:

$$\rho_I(\mathbf{x}, \mathbf{v}) = I_2(\mathbf{W}(\mathbf{x}, \mathbf{v})) - I_1(\mathbf{x}) \quad (8)$$

$$\rho_Z(\mathbf{x}, \mathbf{v}) = Z_2(\mathbf{W}(\mathbf{x}, \mathbf{v})) - \left(Z_1(\mathbf{x}) + \mathbf{D}^T\mathbf{v}\right) \quad (9)$$

with $D^T = (0, 0, 1)$.

Intensity and depth constraints can be put together in a regularization framework, e.g., using TV, to compute a dense scene flow. Although the scene is parametrized in 2D, the regularization is performed for each component of the 3D motion field, favoring rigid motions. This 2D regularization is more reliable than encouraging local smoothness on apparent motion, as in optical flow methods.

Most scenes of interest can be well modeled as locally rigid scenes, i.e. they can be seen as scenes composed by independent 3D rigid parts. We use this assumption to state the data term as a fidelity measure of an estimated local scene flow for each image pixel. Accordingly, we consider that all pixels on each image neighborhood belong to the same rigid surface in 3D. We do not impose any constraint on this small surface but we assume that between consecutive frames it only performs a translational motion, i.e., the rotation component is negligible. Thus, data term is now expressed as

$$E_D(\mathbf{v}) = \sum_{\mathbf{x}} \sum_{\mathbf{x}' \in N(\mathbf{x})} \Psi\left(\left|\rho_I\left(\mathbf{x}', \mathbf{v}(\mathbf{x})\right)\right|^2\right) + \lambda \Psi\left(\left|\rho_Z\left(\mathbf{x}', \mathbf{v}(\mathbf{x})\right)\right|^2\right), \quad (10)$$

where terms $\rho_I(\mathbf{x}, \mathbf{v})$ and $\rho_Z(\mathbf{x}, \mathbf{v})$ measure the consistency of the local scene flow on intensity and depth data, respectively. This penalty function allows to compute each local scene flow vector as a reweighted least squares problem, reducing the effect of outliers.

4.2. Sparse matching term

The solution of the scene flow is based on a linearization of the data term, so that $\mathbf{v}$ is required to be small in order to the regularization holds. To deal with large motions the solution must be performed in a coarse-to-fine warping procedure where the motion of larger structures is used as initial guess. If the motion of smaller structures is similar to the motion of larger structures, the coarse-to-fine approach works well and improves global convergence. However, if the flow of a structure is smaller than its displacement and its motion is no coherent with larger structures, the scene flow is not well estimated. Larger structures dominate the estimation since local minima in higher scales prevent the correct estimation.

We include a constraint to enforce the consistency of the scene flow with a set of sparse 2D measures computed using SURF features. In addition, depth changes on the set of positions is used to

penalize deviations of the Z component of the estimated scene flow. Features detection and matching is done at the resolution level on the intensity images. Let $\left\{\left(\mathbf{x}_1^1, \mathbf{x}_2^1\right), ..., \left(\mathbf{x}_1^N, \mathbf{x}_2^N\right)\right\}$ be the set of correspondences obtained by descriptor matching. To indicate the corresponding match in frame 2 of some point $\mathbf{x}$ in frame 1, we define the matching function $m(\mathbf{x})$ as follows

$$m(\mathbf{x}) = \begin{cases} \mathbf{x}_2^i & \text{if } \left|\mathbf{x} - \mathbf{x}_1^i\right| < \mu_M \\ 0 & \text{otherwise} \end{cases}, \tag{11}$$

where the parameter μ_M is used to control the influence of each matched point on its neighborhood. The matching term is defined as

$$E_M(\mathbf{v}) = \sum_{\mathbf{x}} p(\mathbf{x}) \Psi\left(\left|\delta_{3D}\left(\mathbf{x}, m(\mathbf{x})\right) - \mathbf{v}(\mathbf{x})\right|^2\right) \tag{12}$$

whit $p(\mathbf{x}) = 1$ if there is a descriptor in the interest region around point $\mathbf{x}$. To measure the consistency of the scene flow and the set of matched points, the function $\delta_{3D}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{M}_{\text{cam}}^{-1}(\mathbf{x}_2 Z_2(\mathbf{x}_2) - \mathbf{x}_1 Z_1(\mathbf{x}_1))$ computes the 3D displacement for each correspondency. The robust norm Ψ is applied to deal with wrong matches.

4.3. Regularization term

TV regularization has proved to be very effective in optical flow methods. We use a TV regularization of the 3D motion field to favor locally rigid motions and preserve motion discontinuities. Since we are provided with reliable information of the surface we use the depth data to adapt the regularization. In most cases discontinuities of the 3D motion field coincide with the boundary of the observed 3D surface. As the depth image $Z(\mathbf{x})$ is a 2D parametrization of the 3D surface, we use $|\nabla Z(\mathbf{x})|$ to quantify the discontinuities of the surface. We define the decreasing positive function

$$\omega(\mathbf{x}) = \exp\left(-\alpha |\nabla Z_1(\mathbf{x})|^\beta\right) \tag{13}$$

to prevent regularization of the motion field along strong depth discontinuities. The regularization term is given by:

$$E_R(\mathbf{v}) = \sum_{\mathbf{x}} \omega(\mathbf{x}) \left|\nabla \mathbf{v}(\mathbf{x})\right|, \tag{14}$$

where we use the notation $|\nabla \mathbf{v}| := |\nabla v_X| + |\nabla v_Y| + |\nabla v_Z|$.

5. OPTIMIZATION

To compute the scene flow we introduce in (6) an auxiliary flow $\mathbf{u}$ (following [15]) and solve for the 3D motion field $\mathbf{v}$ that minimizes

$$E(\mathbf{v}, \mathbf{u}) = E_D(\mathbf{v}) + \alpha E_M(\mathbf{v}) + \frac{1}{2\theta}\left|\mathbf{v} - \mathbf{u}\right|^2 + \beta E_R(\mathbf{u}) \tag{15}$$

where θ is a small constant. It can been observed that (15) approaches (6) if $\theta \to 0$. The auxiliary flow $\mathbf{u}$ allows to decompose the optimization into two simpler problems, which can be solved by alternating the updating of $\mathbf{u}$ and $\mathbf{v}$.

1. For a fixed $\mathbf{v}$, we solve for $\mathbf{u}$ that minimizes

$$\sum_{\mathbf{x}} \frac{1}{2\kappa}\left|\mathbf{u}(\mathbf{x}) - \mathbf{v}(\mathbf{x})\right|^2 + \omega(\mathbf{x})\left|\nabla \mathbf{u}(\mathbf{x})\right| \tag{16}$$

where $\kappa = \beta\theta$. For every dimension this problem corresponds to a weighted version of the TV formulation for image denoising. An efficiently solution is given by the dual-ROF model [14] as follows

$$u_d(\mathbf{x}) = v_d(\mathbf{x}) + \kappa\, \omega(\mathbf{x}) \left(\operatorname{div} \mathbf{p}\right)(\mathbf{x})$$

for each dimension $d = X, Y, Z$. The dual variable $\mathbf{p} = \frac{\nabla u_d}{|\nabla u_d|}$ is defined recursively by

$$\mathbf{q}^{n+1}(\mathbf{x}) = \mathbf{p}^n(\mathbf{x}) + \frac{\tau}{\kappa}\left(\nabla u_d^{n+1}\right)(\mathbf{x})$$
$$\mathbf{p}^{n+1}(\mathbf{x}) = \frac{\mathbf{q}^{n+1}(\mathbf{x})}{\max\left\{1, |\mathbf{q}^{n+1}(\mathbf{x})|\right\}}$$

where $\mathbf{p}^0 = 0$ and $\tau \leq 1/4$.

2. For a fixed $\mathbf{u}$, we solve for $\mathbf{v}$ that minimizes

$$E_D(\mathbf{v}) + \alpha E_M(\mathbf{v}) + \sum_{\mathbf{x}} \frac{1}{2\theta}\left|\mathbf{v}(\mathbf{x}) - \mathbf{u}(\mathbf{x})\right|^2 \tag{17}$$

Considering that an initial estimate $\mathbf{v}$ is known, we solve iteratively for increments $\Delta\mathbf{v} = (\Delta v_X, \Delta v_Y, \Delta v_Z)^T$ Using a first-order Taylor series expansion on (17) yields to

$$\begin{aligned} \sum_{\mathbf{x}} \sum_{\mathbf{x}' \in N(\mathbf{x})} & \Psi\left(\left|\rho_I\left(\mathbf{x}', \mathbf{v}(\mathbf{x})\right) + \left(\nabla_I \mathbf{J}\right)\Delta\mathbf{v}(\mathbf{x})\right|^2\right) \\ & + \lambda \Psi\left(\left|\rho_Z\left(\mathbf{x}', \mathbf{v}(\mathbf{x})\right) + \left(\nabla_Z \mathbf{J}\right)\Delta\mathbf{v}(\mathbf{x}) - \mathbf{D}\right|^2\right) \\ & + \alpha\, p(\mathbf{x}) \Psi\left(\left|\delta_{3D}\left(\mathbf{x}, m(\mathbf{x})\right) - \left(\mathbf{v} + \Delta\mathbf{v}\right)(\mathbf{x})\right|^2\right) \\ & + \frac{1}{2\theta}\left|\mathbf{u}(\mathbf{x}) - \left(\mathbf{v} + \Delta\mathbf{v}\right)(\mathbf{x})\right|^2 \end{aligned} \tag{18}$$

where $\nabla_I = (\partial_x I, \partial_y I)$ and $\nabla_Z = (\partial_x Z, \partial_y Z)$, both evaluated at $\mathbf{W}(\mathbf{x}; \mathbf{v})$, and with $\mathbf{J} = \frac{\partial \mathbf{W}}{\partial \mathbf{v}}$ the Jacobian of the warp function. This optimisation problem can be solved independently for every $\mathbf{x}$ using IRLS. Let $\Psi'(s^2)$ denote the derivative of Ψ with respect s^2, using the auxiliary flow $\mathbf{u}$ as an initial estimate of $\mathbf{v}$, the scene flow increment can be computed by

$$\begin{aligned} \Delta\mathbf{v} = \mathbf{H}^{-1} \sum_{\mathbf{x}' \in N(\mathbf{x})} & \Big\{-\Psi'\left(\rho_I^2\left(\mathbf{x}', \mathbf{v}\right)\right)\left(\nabla_I \mathbf{J}\right)^T \rho_I\left(\mathbf{x}', \mathbf{v}\right) \\ & -\lambda\, \Psi'\left(\rho_Z^2\left(\mathbf{x}', \mathbf{v}\right)\right)\left(\nabla_Z \mathbf{J} - \mathbf{D}\right)^T \rho_Z\left(\mathbf{x}', \mathbf{v}\right)\Big\} \\ & + \alpha\, p(\mathbf{x}) \Psi'\left(\rho_{3D}^2\left(\mathbf{x}, \mathbf{v}\right)\right) \rho_{3D}\left(\mathbf{x}, \mathbf{v}\right) + \frac{1}{2\theta}(\mathbf{u} - \mathbf{v}) \end{aligned} \tag{19}$$

where ρ_{3D} is a 3D residual defined as

$$\rho_{3D}\left(\mathbf{x}, \mathbf{v}\right) = \delta_{3D}\left(\mathbf{x}, m(\mathbf{x})\right) - \mathbf{v}, \tag{20}$$

and $\mathbf{H}$ is the Gauss-Newton approximation of the Hessian matrix. Unlike [13], the proposed regularization ensures $\mathbf{H}$ to be nonsingular allowing to estimate the 3D motion even on untextured regions.

6. EXPERIMENTS

In order to validate our local/global method (LG_{SF}) we use the Middlebury stereo database [16]. Using images of these datasets is equivalent to a fixed camera observing a moving object along X axis. We take image 2 of each dataset as the first frame and image 6 as the second frame, both in quarter-size. The ground truth for the optical flow is given by the disparity map from frame 1.

We compare our approach with other scene flow methods: stereo based approaches [8] and [9], particle filtering based method (PF_{SF}) [3], locally rigid approach (L_{SF}) [13] and an orthographic variation of LG_{SF} which is denoted by ORT_{SF}. We also include results for the scene flow inferred using TV-L^1 optical flow [15] and depth data. Results are computed considering all non-occluded pixels.

To compare methods we use 2 measures: the *normalized root mean squared* of the optical flow difference (NRMS_{OF}) and the

	Teddy		*Cones*	
	$\mathrm{NRMS_{OF}}$	AAE	$\mathrm{NRMS_{OF}}$	AAE
$\mathrm{LG_{SF}}$	0.0222	0.837	0.0164	0.526
TV-L^1	0.0642	1.360	0.0509	0.932
$\mathrm{L_{SF}}$	0.0780	2.288	0.0577	1.991
$\mathrm{ORT_{SF}}$	0.0811	0.866	0.0594	0.963
[9]	0.0285	1.010	0.0307	0.390
[8]	0.0621	0.510	0.0579	0.690
$\mathrm{PF_{SF}}$	0.110	5.040	0.090	5.020

Table 1. Optical flow errors for Middlebury dataset.

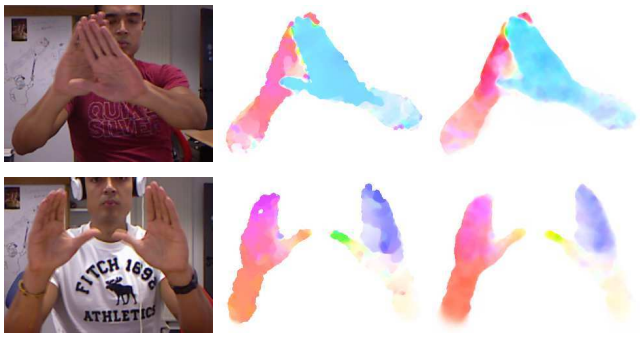
	Original		*Modified*	
	$\mathrm{NRMS_{SF}}$	P10%	$\mathrm{NRMS_{SF}}$	P10%
$\mathrm{LG_{SF}}$	0.0353	97,55	0.0754	90,28
TV-L^1	0.5493	84,94	0.4662	84,85
$\mathrm{L_{SF}}$	0.4415	89,07	0.3039	83,16
$\mathrm{ORT_{SF}}$	0.4678	82,77	0.4999	82,34

Table 2. Scene flow errors for original and modified datasets.

average angle error (AAE). The $\mathrm{NRMS_{OF}}$ is normalized by the range of the optical flow magnitude of each dataset. In this part, we consider 2D errors since this the only available information for all methods. Results are presented in Table 1. $\mathrm{LG_{SF}}$ achieves the best $\mathrm{NRMS_{OF}}$ results and its AAE is comparable with stereo based methods. On the other hand, in comparison with $\mathrm{PF_{SF}}$ and $\mathrm{L_{SF}}$, which also use intensity and depth, the improvement is notable. A non-optimized version of $\mathrm{LG_{SF}}$ processed the dataset, on a double core desktop machine in under 10 s, as opposed to 5 h for [8] or 10 min for [3] (run time was not reported in [9]).

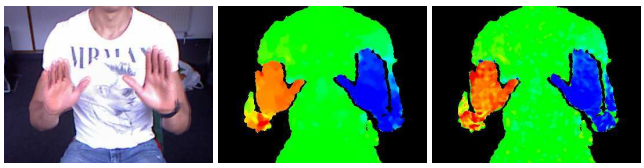
In order to asses scene flow errors we compute the *normalized root mean squared* ($\mathrm{NRMS_{SF}}$) of the 3D motion difference. Besides, we compute the statistic P10% of $\mathrm{NRMS_{SF}}$ to show the percentage of pixels with a scene flow estimation within 10% of the ground truth magnitude. The ground truth for the scene flow is constant and given by the baseline of the stereo setup. To consider a more challenging experiment we modified the original dataset to include a virtual Z-motion by scaling the frame 2 with a scale factor $S > 1$. In this way the ground truth for the scene flow is not constant anymore and the Z-motion of a point with depth Z is given by $-(1-S)Z$. Average results for images Teddy and Cones are shown in Table 2. Once again $\mathrm{L_{SF}}$ presents the best performance.

Finally, some experiments were performed from image sequences of a Microsoft Kinect sensor. In Figure 1 we show the results of the motion field estimation between a pair of images with hand-hand overlapping and another performing a subtle motion of both hands with rotation. For the computation only points less than 2 meters away from the sensor are considered. Figure 2 shows results for the Z motion estimation in a sequence where the left (in the image) hand is moving away from the sensor while the right one is approaching it. We also presents the results of the motion estimation by TV-L^1 where the scene flow is inferred using the depth data. Both for the 2D motion field and the Z motion, $\mathrm{LG_{SF}}$ allows a more reliable motion estimation, even over low textured regions. Moreover, the application of a 2D TV on the components of the 3D motion field overcomes the over-regularization (as in the TV-L^1 case), while preserving motion and structure discontinuities.



(a) Input frames (b) $\mathrm{LG_{SF}}$ (c) TV-L^1

Fig. 1. 2D motion field estimation. Results are presented using the Middlebury color code [16].



(a) Input frames (b) $\mathrm{LG_{SF}}$ (c) TV-L^1

Fig. 2. Z motion estimation. Results are presented using a cold-to-warm color map for the range [-2.5,2.5] cm, where the zero velocity is green, blue and magenta colors represent negative velocities (approaching pixels) and red and yellow colors are positives velocities.

7. CONCLUSIONS

We proposed a novel approach to compute a dense scene flow using intensity and depth data. We combine local and global constraints to solve for the 3D motion field in a variational framework. Unlike previous intensity and depth-based methods, in this work depth data is used in 3 ways: to model the motion in the image domain, to constrain the scene flow and to adapt the TV regularization. Assuming a local rigidity of the scene, our method promotes local constancy of the motion and the dense solution is achieved by alternating this procedure with an adaptive 2D TV regularization of each component of the 3D motion field. This approach allows to solve for a dense scene flow while preserving motion discontinuities.

Throughout some experiments, we demonstrated the validity of our method. The local/global approach outperforms previous scene flow methods in the optical flow estimation for the Middlebury stereo dataset. Moreover, the proposed approach is many times faster. Because of the lack of a proper benchmark, we modified this dataset to test our method in the estimation of a more challenging motion field. In this experiment, the Z component of the scene flow ground truth is not constant anymore and results of $\mathrm{LG_{SF}}$ stands the best. Besides, using data from the Kinect sensor, we estimated scene flow for specific motions and the results are encouraging.

We are currently exploring how to deal with occlusions and optimizing our implementation to achieve real-time processing. Besides, we are investigating how to define scene flow-based descriptors to perform action or gestures recognition.

8. REFERENCES

[1] Li Xu, Jiaya Jia, and Yasuyuki Matsushita, "Motion detail preserving optical flow estimation," *IEEE Transaction on Pattern Analysis and Machine Intelligence.*, vol. 34, no. 9, pp. 1744–1757, 2012.

[2] Andreas Wedel, Thomas Pock, Christopher Zach, Horst Bischof, and Daniel Cremers, "An improved algorithm for TV-L1 optical flow," *Statistical and Geometrical Approaches to Visual Motion Analysis*, pp. 23–45, 2009.

[3] S. Hadfield and R. Bowden, "Kinecting the dots: Particle based scene flow form depth sensors," in *International Conference on Computer Vision*, 2011.

[4] J. Quiroga, F. Devernay, and J. Crowley, "Scene flow by tracking in intensity and depth data," in *Conference on Computer Vision and Pattern Recognition Workshops*, 2012.

[5] S. Vedula, S. Baker, P. Rander, and R. Collins, "Three-dimensional scene flow," in *International Conference on Computer Vision*, 1999.

[6] S. Vedula, S. Baker, P. Render, R. Collins, and T. Kanade, "Three-dimensional scene flow," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 275–280, 2005.

[7] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke, and D. Cremers, "Efficient dense scene flow from sparse of dense stereo data," in *European Conference on Computer Vision*, 2008.

[8] F. Huguet and F. Devernay, "A variational method for scene flow estimation from stereo sequences," in *International Conference on Computer Vision*, 2007.

[9] T. Basha, Y. Moses, and N. Kiryati, "Multi-view scene flow estimation: A view centered variational approach," in *Conference on Computer Vision and Pattern Recognition*, 2010.

[10] C. Vogel, K. Schindler, and S. Roth, "3D scene flow estimation with a rigid motion prior," in *International Conference on Computer Vision*, 2011.

[11] H. Spies, B. Jahne, and J. Barron, "Dense range flow from depth and intensity data," in *International Conference on Pattern Recognition*, 2000.

[12] T. Lukins and R. Fisher, "Colour constrained 4D flow," in *British Machine Vision Conference*, 2005.

[13] J. Quiroga, F. Devernay, and J. Crowley, "Local scene flow by tracking in intensity and depth," *Journal of Visual Communication and Image Representation*, 2013. doi: 10.1016/j.jvcir.2013.03.018.

[14] Antonin Chambolle, "Total variation minimization and a class of binary mrf models," *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 136–152, 2005.

[15] Christopher Zach, Thomas Pock, and Horst Bischof, "A duality based approach for realtime TV-L1 optical flow," in *DAGM-Symposium*, 2007, pp. 214–223.

[16] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Conference on Computer Vision and Pattern Recognition*, 2003.