Joint Large Displacement Scene Flow and Occlusion Variational Estimation

Roberto P. Palomares, Gloria Haro and Coloma Ballester
Universitat Pompeu Fabra, Barcelona, Spain

Keywords: Scene Flow, Variational Methods, Coordinate Descent, Sparse Matches.

Abstract: This paper presents a novel variational approach for the joint estimation of scene flow and occlusions. Our method does not assume that a depth sensor is available. Instead, we use a stereo sequence and exploit the fact that points that are occluded in time, might be visible from the other view and thus the 3D geometry can be densely reinforced in an appropriate manner through a simultaneous motion occlusion characterization. Moreover, large displacements are correctly captured thanks to an optimization strategy that uses a set of sparse image correspondences to guide the minimization process. We include qualitative and quantitative experimental results on several datasets illustrating that both proposals help to improve the baseline results.

1 INTRODUCTION

The structure and motion of objects in a 3D space is an important characteristic of dynamic scenes. Measuring the three-dimensional motion vector fields remains one of the unsolved tasks in computer vision although progress has been made in recent years (e.g., (Basha et al., 2013; Jaimez et al., 2015; Quiroga et al., 2014; Sun et al., 2015; Vogel et al., 2015; Menze and Geiger, 2015; Wedel et al., 2011)) and is currently gaining increasing attention. Reliable 3D motion maps may be used in a wide range of applications such as autonomous robot navigation, driver assistance, augmented reality, 3D movie and TV generation, surveillance or tracking, to mention just a few.

The scene flow problem was defined as the estimation of dense 3D geometry and 3D motion field from nonrigid 3D data (Vedula et al., 2005). In the existing methods, the corresponding vector field is computed either from stereo video sequences taken from different points of view or from monocular RGB-Depth sequences, that is, videos recorded with a camera equipped with a depth sensor. We propose a scene flow method for the first kind of data: stereo sequences.

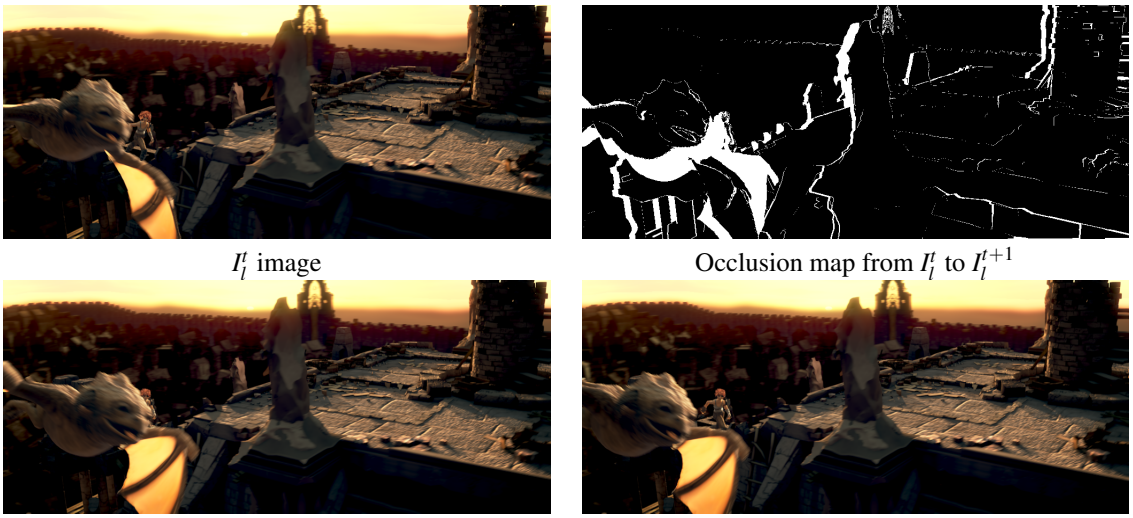
Our contribution in this paper is twofold: we first propose a novel variational approach for the joint estimation of scene flow and motion occlusion; and second, we propose an optimization strategy for variational scene flow which is able to capture large displacements without a multi-scale methodology and is applicable to any scene flow variational method. As for the first contribution, our method uses a sequence of image pairs obtained from two synchronized cameras and simultaneously computes the optical flow between consecutive frames, the corresponding occlusions due to motion and the disparity change between the stereo image pairs. Let us notice that this information, together with calibration data, is an equivalent representation of the 3D scene flow. Regarding our second contribution, we present and show the potential of our general variational scene flow optimization strategy on the proposed energy model which, in turn, has a transparent and generic structure.

The remainder of the paper is organized as follows. In Section 2 we revise previous works on scene flow. Section 3 presents our proposed scene flow energy formulation and the proposed minimization procedure is explained in Section 4. Section 5 presents experimental results. Finally, the conclusions are summarized in Section 6.

2 RELATED WORK

From the seminal work of (Vedula and et al., 1999), several methods have been proposed for the scene flow problem in order to improve the initial formulation which decoupled the computation of 2D optical flow fields and 3D structure. There are mainly two different approaches to face the problem. One of

P. Palomares R., Haro G. and Ballester C.
Joint Large Displacement Scene Flow and Occlusion Variational Estimation.
DOI: 10.5220/0006110601720180
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 172-180
ISBN: 978-989-758-227-1



I_l^t image | Occlusion map from I_l^t to I_l^{t+1}

I_l^{t+1} image | I_r^{t+1} image

Figure 1: Motivation of the proposed data terms and their dependence on the occlusion map. Notice that most part of the girl in I_l^t is not visible in I_l^{t+1} while it is visible in I_r^{t+1}. Thus, the deactivation of the data term between images I_l^t and I_l^{t+1} together with the activation of the data term relating I_l^t and I_r^{t+1} will result in a better estimation of the scene flow variables.

them estimates the scene flow from RGB-Depth data benefiting from the availability of depth data provided by cameras equipped with a depth sensor. The 3D scene flow is estimated directly from it and regularization of the flow field is imposed on the 3D surfaces of the observed scene instead of on the image plane (Pons et al., 2007; Basha et al., 2013; Jaimez et al., 2015; Quiroga et al., 2014; Sun et al., 2015; Vogel et al., 2015). For instance, (Basha et al., 2013; Vogel et al., 2011) jointly estimate depth and a 3D flow field using a variational method which imposes geometric multi-view consistency and 3D smoothness. Some of these methods also use a local rigidity assumption (Menze and Geiger, 2015; Quiroga et al., 2014) representing the dynamic scene, e.g., as a collection of rigidly moving planes (Vogel et al., 2015). The second kind of methods work on stereo video sequences and estimate from them disparity (between the stereo pair) and motion (between consecutive frames) using formulations which mutually constrain the scene flow (Huguet and Devernay, 2007; Wedel et al., 2011). The authors of (Wedel et al., 2011) propose to precompute the stereo disparity and decouple depth and motion estimation by estimating the optical flow and the disparity change through time.

In most of the proposals, the problem is frequently modeled by variational methods where the unknowns representing the motion of each 3D point in the scene are estimated as the minimum of an energy functional (e.g., (Vedula et al., 2005; Pons et al., 2007; Huguet and Devernay, 2007; Basha et al., 2013; Menze and Geiger, 2015; Wedel et al., 2011)). The optimization usually proceeds in a multi-scale or coarse-to-fine procedure and thus smooth motions are favoured and large displacements of small objects are mostly lost.

The variational method we propose does not assume a depth sensor is available nor calibrated cameras. As in (Huguet and Devernay, 2007; Wedel et al., 2011), we use a two-view setup with a pair of stereo image frames. Our proposal also estimates motion occlusions and benefits from the appropriate comparison among views of the scene. In order to correctly estimate large displacements of small objects, our minimization works by incorporating sparse matches which drive the minimization of the energy in local patches, providing a fast method that works at the finest scale, i.e., the original scale of the image data.

3 SCENE FLOW MODEL

Let us assume that a stereo video sequence is given, consisting of different image pairs that have been obtained from two views. For each time instant t, let $I_l^t, I_r^t, I_l^{t+1}, I_r^{t+1} : \Omega \to \mathbb{R}$ be two of those consecutive stereo pair of frames of the stereo video sequence, where the subscripts l and r stand for *left* and *right*, respectively, and t stands for time. As usual, we assume that the image domain Ω is a rectangle in $\mathbb{R}^2$. Our starting point will be the model for scene flow introduced in (Wedel et al., 2011), where a decoupled approach was presented. In a decoupled approach,

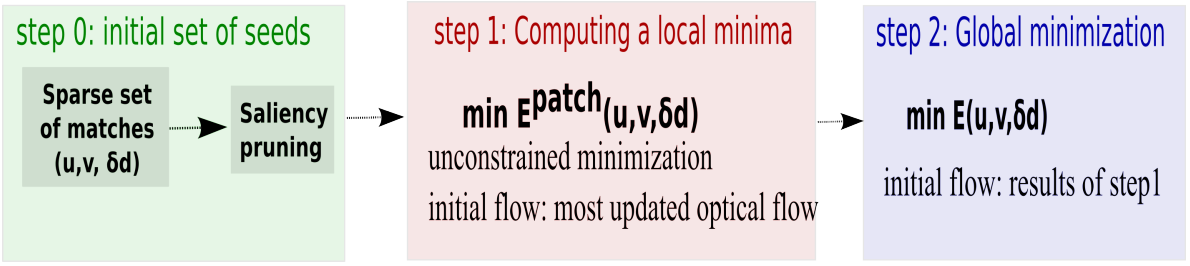


Figure 2: Diagram with the main steps of the proposed method.

the estimation of depth or disparity at fixed time is done previously to, and independently of, the estimation of the motion (optical flow and disparity change). This problem separation provides more flexibility and has some advantages as the disparity may be estimated with an optimal stereo algorithm. The decoupled scene flow approach enforces a coupling among disparity, optical flow, and disparity change.

Let d be a given disparity map between I_l^t and I_r^t. Let $\mathbf{u} = (u, v)$ denote the optical flow between the left frames, I_l^t and I_l^{t+1}, and δd denote the change in disparity between the stereo pairs at times t and $t+1$. In order to write the energy model in a more compact form, let us first introduce the following notation:

$$\begin{aligned}
D_1 &= I_l^{t+1}(x+u, y+v) - I_l^t(x, y) \\
D_2 &= I_l^{t+1}(x+u, y+v) - I_r^t(x+d, y) \\
D_3 &= I_r^{t+1}(x+d+u+\delta d, y+v) - I_l^{t+1}(x+u, y+v) \\
D_4 &= I_r^{t+1}(x+d+u+\delta d, y+v) - I_l^t(x, y) \\
D_5 &= I_r^{t+1}(x+d+u+\delta d, y+v) - I_r^t(x+d, y)
\end{aligned}$$

In order to compute the scene flow field $(u, v, \delta d)$, Wedel et al. (Wedel et al., 2011) propose to minimize an energy functional which is made of two terms, namely, $\bar{E}(u, v, \delta d) = \bar{E}_R(u, v, \delta d) + \bar{E}_D(u, v, \delta d)$, where

$$\begin{aligned}
\bar{E}_R(u, v, \delta d) &= \alpha \int_\Omega \Psi(|\nabla u|^2 + |\nabla v|^2 + \gamma |\nabla \delta d|^2) dxdy \\
\bar{E}_D(u, v, \delta d) &= \int_\Omega \Psi(|D_1|^2) dxdy \\
&+ \int_\Omega o \Psi(|D_3|^2) dxdy + \int_\Omega o \Psi(|D_5|^2) dxdy
\end{aligned}$$

where $\Psi(s^2) = \sqrt{s^2 + \varepsilon^2}$, with $\varepsilon = 0.0001$ being a small constant, and $o(x, y, t)$ is the given stereo visibility map for the given disparity map d (i.e., $o(x, y, t) = 1$ if (x, y) is visible both in I_l^t and in I_r^t). We have omitted in $\bar{E}, \bar{E}_R, \bar{E}_D$ the dependency of $u, v, \delta d, d, o$ on x, y, t for the sake of simplicity. Finally, let us notice that the regularity term is based on a differentiable approximation of the Total Variation. Similarly, the data term is based on the same differentiable approximation of the L^1 norm of the constraints favoring constancy in intensity of the same point in the scene, thus in the four involved images.

This method does not directly take occlusions into account and relies on data terms that consider correspondence errors even for the occluded pixels where no correspondence can be established. Hence, erroneous flows are generated at moving occlusion boundaries. Explicitly modeling occlusions has proved beneficial in optical flow estimation methods (e.g. (Ayvaci et al., 2012; Ballester et al., 2012; Ince and Konrad, 2008) among others). Occlusion reasoning has been considered in scene flow estimation methods that use depth sensors (Wang et al., 2015; Zanfir and Sminchisescu, 2015). On the other hand, it is traditionally believed that motion vectors tend to be smaller in magnitude than disparities, especially if the video sequences have been captured with a small time delay; but this assumption does not hold for the current standard databases (Butler et al., 2012; Geiger et al., 2012) which contain important large displacements. In these situations, handling occlusions due to motion is as important as handling occlusions due to disparity.

In this work we extend the previous model to jointly compute the optical flow, its associated occlusions, and the disparity change. Let $\chi : \Omega \to [0, 1]$ be the function modeling the motion occlusion map, so that $\chi(x, y, t) = 1$ identifies the motion occluded pixels, i.e. pixels that are visible in I_l^t but not in I_l^{t+1}. Our model is based on the assumption that the occluded region due to motion, given by $\chi(x, y, t) = 1$, should include the region where the divergence of the optical flow is negative. This was pointed out by Sand and Teller (Sand and Teller, 2008), who noticed that the divergence of the motion field may be used to distinguish between different types of motion areas. Schematically, the divergence of a flow field is negative for occluded areas, positive for disoccluded, and near zero for the matched areas. Taking this into account, Ballester et al. (Ballester et al., 2012) proposed a variational model for the joint estimation of occlu-

sions and optical flow. In order to consider motion occlusions and benefit from an appropriate comparison among the different views of the scene, we build up from these ideas and propose to include a new term in the energy functional that characterizes the occlusion areas as those where the divergence of the flow is negative. We also propose to include different types of data terms in the energy functional which are activated based on the occlusion information provided by χ. In this way, if there is a motion occlusion in the left view, $\chi = 1$, the energy will only consider error correspondences in the right views, where the object is still visible. Figure 1 presents an example motivating our proposal; by detecting the occlusion regions, the motion field in these regions will be recovered by using the fact that they are visible in the remaining views which we use to introduce new data constraints. Thus, the proposed energy contains three parts, namely,

$$E(u,v,\delta d,\chi) = E_R(u,v,\delta d,\chi) + E_D(u,v,\delta d,\chi) + E_{occ}(u,v,\chi) \quad (1)$$

where

$$\begin{aligned}
E_R(u,v,\delta d,\chi) &= \alpha \int_\Omega \Psi\left(|\nabla u|^2+|\nabla v|^2+\gamma|\nabla \delta d|^2\right) \\
&+ \eta \int_\Omega \Psi\left(|\nabla \chi|^2\right) \\
E_{occ}(u,v,\chi) &= \beta \int_\Omega \chi div(u,v) dxdy \\
E_D(u,v,\delta d,\chi) &= \int_\Omega (1-\chi)\, \Psi(|D_1|^2) dxdy \\
&+ \int_\Omega (1-\chi)\, o\, \Psi(|D_2|^2) dxdy \\
&+ \int_\Omega (1-\chi)\, o\, \Psi(|D_3|^2) dxdy \\
&+ \int_\Omega \chi\, o\, \Psi(|D_4|^2) dxdy \\
&+ \int_\Omega \chi\, o\, \Psi(|D_5|^2) dxdy
\end{aligned}$$

Again, the map χ is evaluated in (x,y,t) in the functional but we omit it for the ease of notation.

4 OPTIMIZATION STRATEGY

In order to make the optimization problem more tractable, optical flow variational methods include a linearization of the warped images in the data terms, which leads to embed the functional into a coarse-to-fine multi-level approach to better handle large motion fields. However, this approach still fails to recover large motions of small objects not present at coarser scales. Different approaches to overcome this limitation have been proposed in the past years. Among the most recent works, several ones share the trait of being based on a sparse-to-dense estimation that avoids the classical coarse-to-fine scheme. They start with a set of correspondences (non-dense feature-based matches), which are used to generate a dense optical flow field and subsequently, the next step produces a global refinement over the whole image domain. For instance, in the work (Palomares et al., 2016), an initial set of sparse matches is grown by a coordinate descent scheme used to minimize the target energy functional. Our proposal builds upon these ideas to propose a minimization method for the scene flow energy. Figure 2 shows a diagram with the main steps of the proposed algorithm. The optimization process works in two stages, with a previous initialization of the sparse matches (named as zero stage in the following), both of them operating at the finest scale of the image:

0. The zero stage builds the initial set of sparse *seeds* $(u,v,\delta d)$. The algorithm assumes that a set of sparse correspondences between two pairs of images are provided; in particular, between $I_l^t \leftrightarrow I_l^{t+1}$ and $I_l^{t+1} \leftrightarrow I_r^{t+1}$. In order to estimate sparse correspondences between both pairs of images we use the DeepMatching algorithm (Weinzaepfel et al., 2013). From the first set of matchings, between $I_l^t \leftrightarrow I_l^{t+1}$, we obtain an initial set of candidates for the variables (u,v). Then, to completely define the set of seeds for solving the scene flow problem, it is necessary to find an estimation of $\delta d(\mathbf{x})$ associated to each optical flow candidate $(u(\mathbf{x}),v(\mathbf{x}))$ at the different sparse locations $\mathbf{x} = (x,y) \in \Omega$. From the second set of sparse matches, between $I_l^{t+1} \leftrightarrow I_r^{t+1}$, we select the discrete value of $\hat{d}_{t+1}$ to be the disparity associated of the closest keypoint $\hat{\mathbf{x}}$ in I_l^{t+1} (with a matching in I_r^{t+1}) to the position $(x+u(\mathbf{x}),y+v(\mathbf{x}))$ within a certain tolerated distance. If there exists such a keypoint in I_l^{t+1}, we add $(u(\mathbf{x}),v(\mathbf{x}),\delta d(\mathbf{x}),\chi(\mathbf{x}))$ as an initial seed, where $\delta d(\mathbf{x}) = \hat{d}_{t+1}(\hat{\mathbf{x}}) - d_t(\mathbf{x})$ and $\chi(\mathbf{x}) = 0$.
1. The first stage consists in computing a dense scene flow estimation providing a *good* local minimum of the target energy (the proposed (1) in this paper); good in the sense that captures large displacements and controls the error on occlusion areas. Our method proceeds by minimizing the energy over local neighborhoods (patches) in a proper order defined by the reliability of the scene flow estimation at the center of each patch. This ordering is managed by a priority queue where the most reliable estimations – the estimated $(u,v,\delta d,\chi)$ values that have the lowest energy values – are placed at the top positions of

the queue. Initially, the queue is formed by the sparse set of seeds. These seeds have an associated local energy equal to zero (full reliability). Then, an iterative process is launched; the following procedure is iterated until the priority queue is emptied:

- The top element of the queue of scene flow candidates is extracted and its associated scene flow value is set as visited at its corresponding position.
- The patch around the visited position is considered and a scene flow is interpolated within the patch by propagating the already visited values.
- The scene flow energy is minimized in the patch, starting with the previous interpolation as initialization. Notice that this step can be thought as a minimization of the energy where all the variables outside the patch under consideration have been fixed, thus bearing similarities with the coordinate descent methods.
- The local energy in the patch is computed and the four immediate neighbors of the center pixel are introduced as new candidates in the queue with a reliability given by the local energy (the energy of the patch).

2. The result of the first step, the data corresponding to $(u,v,\delta d,\chi)$, is a dense scene flow estimation providing a good local minimum of the energy (1).This result is refined in the second stage by the minimization of the energy functional over the whole image domain. In other words, the result of the first step is used as an initialization for minimizing the energy around it.

Let us remark that the method of Cech et al. (Cech et al., 2011) also uses an algorithm to estimate both disparity and optical flow from a stereo sequence by growing a set of seeds. In contrast to our seed growing method driven by the energy minimization, the method in (Cech et al., 2011) constructs heuristics based on photometric consistency through correlations and constant parameters adjusting the amount of optical flow regularization and temporal consistency. Moreover, it provides a semi-dense scene flow while we get a dense estimation.

In order to minimize our energy formulation (1), the associated Euler-Lagrange equations are numerically solved. To simplify the presentation, we introduce the following notations

$$\begin{aligned} R_m &= \sqrt{|\nabla u|^2+|\nabla v|^2+\gamma|\nabla \delta d|^2} \\ R_o &= |\nabla \chi| \\ \Psi'(s^2) &= \frac{1}{2\sqrt{s^2+\varepsilon^2}} \end{aligned}$$

Thereby, the Euler-Lagrange equations are

$$\begin{aligned} 0 &= -\alpha\, div\left(\Psi'(R_m^2)\cdot\nabla u\right) \\ &+ (1-\chi)\cdot\Psi'(D_1^2)\cdot D_1\cdot I_{l,x}^{t+1}(x+u,y+v) \\ &+ o\,(1-\chi)\cdot\Psi'(D_2^2)\cdot D_2\cdot I_{l,x}^{t+1}(x+u,y+v) \\ &+ o\,(1-\chi)\cdot\Psi'(D_3^2)\cdot D_3\cdot (I_{r,x}^{t+1}(x+d+u+\delta d,y+v) \\ &\quad -I_{l,x}^{t+1}(x+u,y+v)) \\ &+ o\,\chi\cdot\Psi'(D_4^2)\cdot D_4\cdot I_{r,x}^{t+1}(x+d+u+\delta d,y+v) \\ &+ o\,\chi\cdot\Psi'(D_5^2)\cdot D_5\cdot I_{r,x}^{t+1}(x+d+u+\delta d,y+v) \\ &\quad -\beta\chi_x\ , \\ 0 &= -\alpha\, div\left(\Psi'(R_m^2)\cdot\nabla v\right) \\ &+ (1-\chi)\cdot\Psi'(D_1^2)\cdot D_1\cdot I_{l,y}^{t+1}(x+u,y+v) \\ &+ o\,(1-\chi)\cdot\Psi'(D_2^2)\cdot D_2\cdot I_{l,y}^{t+1}(x+u,y+v) \\ &+ o\,(1-\chi)\cdot\Psi'(D_3^2)\cdot D_3\cdot (I_{r,x}^{t+1}(y+d+u+\delta d,y+v) \\ &\quad -I_{l,y}^{t+1}(x+u,y+v)) \\ &+ o\,\chi\cdot\Psi'(D_4^2)\cdot D_4\cdot I_{r,y}^{t+1}(x+d+u+\delta d,y+v) \\ &+ o\,\chi\cdot\Psi'(D_5^2)\cdot D_5\cdot I_{r,y}^{t+1}(x+d+u+\delta d,y+v) \\ &- \beta\chi_y\ , \\ 0 &= -\alpha\gamma\, div\left(\Psi'(R_m^2)\cdot\nabla\delta d\right) \\ &+ o\,(1-\chi)\Psi'(D_3^2)\cdot D_3\cdot I_{r,x}^{t+1}(x+d+u+\delta d,y+v) \\ &+ o\,\chi\cdot\Psi'(D_4^2)\cdot D_4\cdot I_{r,x}^{t+1}(x+d+u+\delta d,y+v) \\ &+ o\,\chi\cdot\Psi'(D_5^2)\cdot D_5\cdot I_{r,x}^{t+1}(x+d+u+\delta d,y+v)\ , \\ 0 &= -\alpha\eta\, div\left(\Psi'(R_o^2)\cdot\nabla\chi\right) \\ &- \Psi(D_1^2)-\Psi(D_2^2)-o\,\left(\Psi(D_3^2)+\Psi(D_4^2)+\Psi(D_5^2)\right) \\ &+ \beta div(u,v)\ . \end{aligned}$$

where the subindices x and y denote the partial derivatives with respect to x and y, respectively, and the point coordinates (x,y) have been omitted in the gradient expressions.

The Euler-Lagrange equations are non-linear in the unknowns $(u,v,\delta d,\chi)$ due to the multiple warped images $I_l^{t+1}(x+u,y+v), I_{l,x}^{t+1}(x+u,y+v)$, etc. To numerically solve them, either over a local patch or over the whole domain, we follow the optimization method proposed by Brox et al. (Brox et al., 2004). It is based on two fixed iterations loops to cope with the non-linear terms. The external loop is used to handle the linearization of the data terms in the warped form, and the internal loop takes into account the non-linearities of the Ψ' functions. After the linearization, the resulting linear system can be efficiently solved using the SOR method (Young, 1971).

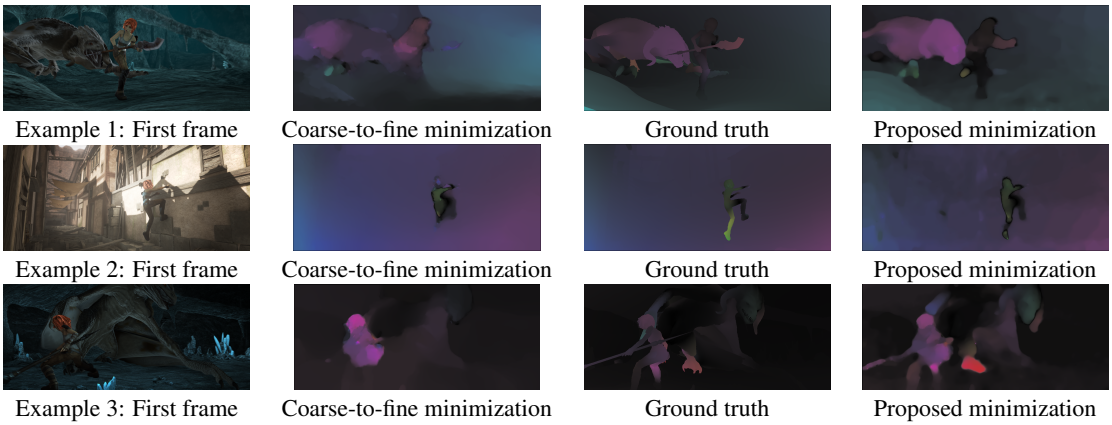


Figure 3: Comparison of two minimization strategies for the same energy proposed in (Wedel et al., 2011). Our minimization startegy is not based on a coarse-to-fine scheme but on sparse correspondences that allow to capture large displacements. Results in the MPI Sintel training set.

5 EXPERIMENTS

In this section we provide two sets of experiments. The first one is designed to validate the better behaviour of the selected optimization strategy against the classic coarse-to-fine multi-level approach. The second one shows the properties of the presented functional due to the explicitly occlusion handling. Let us remark that all results have been obtained by using the grayscale versions of the original color frames. The color version is only used to compute the seeds with the Deep Matching algorithm. The experiments use stereo sequences from the MPI Sintel Flow dataset (Butler et al., 2012) and from the KITTI 2015 dataset (Menze and Geiger, 2015). Sintel has 23 training sequences. For every frame, there are two different versions of the images, “clean” and “final”. The difference is that the second set adds complexity to the first one by incorporating atmospheric effects, depth of field blur, motion blur, color correction and other details. It contains several sequences with large motions of small objects. KITTI contains different sequences of a city provided by an autonomous driving platform. It presents large deformations, dynamic scenes and challenging iluminatons changes.

5.1 Benefits of the New Optimization Strategy

Our first goal is to validate the good performance of the optimization scheme and show the benefits in the presence of large displacement motions against the coarse-to-fine strategy. For this purpose, we use the energy functional proposed by Wedel et al. (Wedel et al., 2011) (detailed at the beginning of Section 3), and we compute the motion field using these two different minimization approaches. In Tables 1 and 2, they are denoted by *Classic Wedel* (i.e., classic coarse-to-fine strategy for the Wedel et al. (Wedel et al., 2011) energy) and *Our Wedel*. Fig. 3, Tables 1 and 2 show that the chosen minimization approach is able to recover large motions where the coarse-to-fine strategy fails. For each group of four images in Fig. 3, from top to bottom and from left to right, the first frame is displayed in (a), the optical flow estimation from the classic coarse-to-fine Wedel et al. (Wedel et al., 2011) is displayed in (b), (c) shows the optical flow ground truth, and (d) the optical flow estimated with our scene flow minimization strategy for the same energy. Let us notice from this figure and also from Tables 1 and 2 that the optimization scheme is also better for the kind of sequences where the multi-scale approach does not fail. It is clear that the integration of sparse matches results with an appropriate minimization strategy directly at the finest image scale represents a great improvement in comparison to the coarse-to-fine optimization strategy.

5.2 Benefits of the New Energy

The second goal is to show the advantages of the proposed energy functional which includes new data terms and motion occlusion estimation. Fig. 4 displays results on three sequences of the MPI Sintel training set. For each group of six images, from top to bottom and from left to right, the first frame is shown in (a), the ground truth occlusions are displayed in (b), (c) shows the optical flow from our scene flow minimization strategy for the Wedel et al. (Wedel et al., 2011) energy (our Wedel), (d) the optical flow ground truth, and (e) and (f) the occlusions and opti-

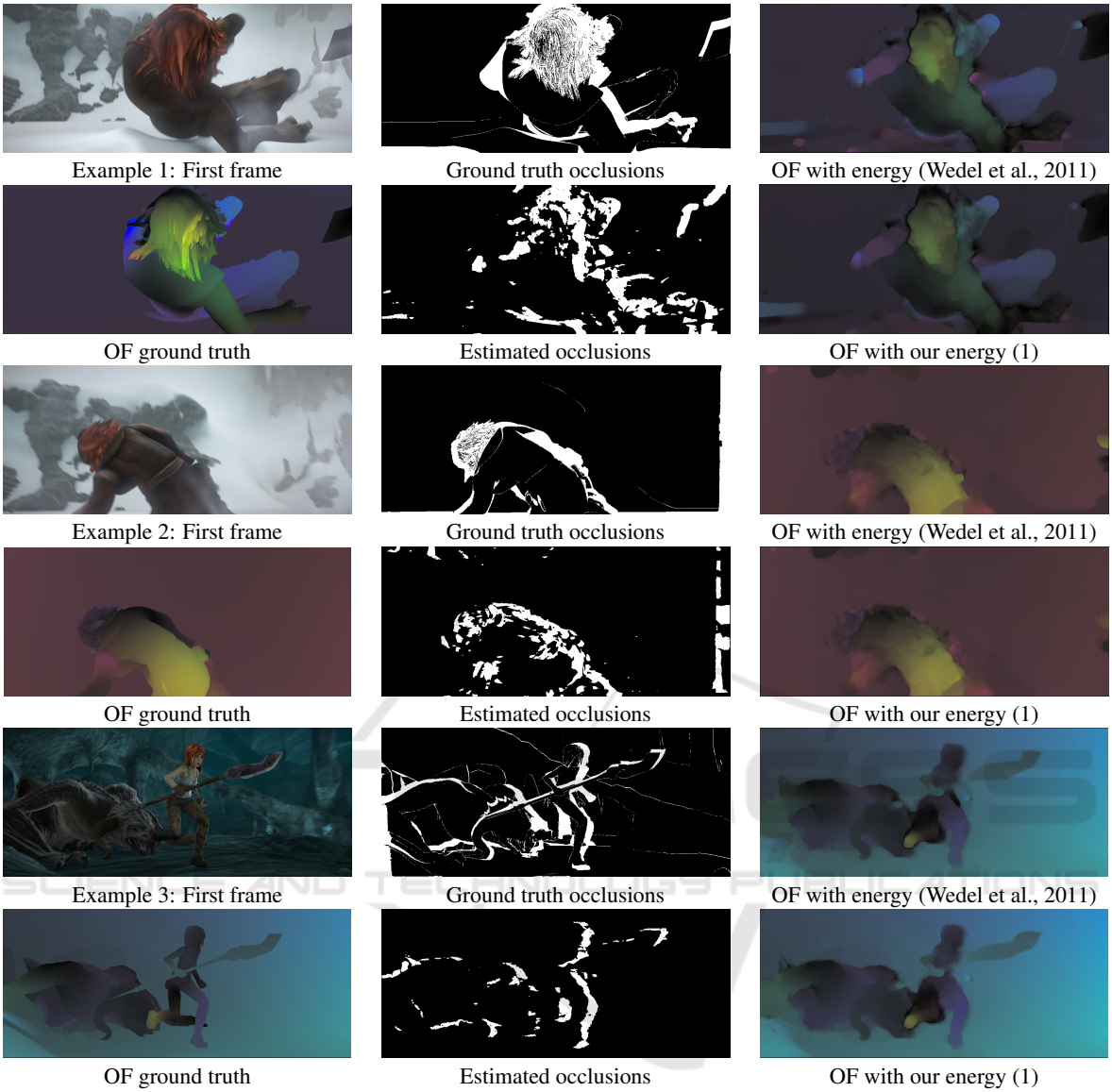


Figure 4: Comparison of the estimated optical flows (OF) estimated with the baseline energy (Wedel et al., 2011) and with the new proposed energy (1) (in both cases using our proposed minimization strategy). Results in the MPI Sintel training set.

cal flow estimated from our whole proposal with the energy (1). On the other hand, the second and third rows of Table 1 and Table 2 show the global accuracy over the whole datasets of both proposals. The results of Table 1 and Fig. 4, show that the proposed energy keeps better results at the visible areas (second column of Table 1) and it specially improves the accuracy at the occluded areas (third column). The occlusion mask allows to densely reinforce the 3D geometry from the fact that points that are occluded in time in the left view might be visible from the other (right) view and thus obtain better optical flow boundaries especially near occlusion regions. This effect is noticeable for instance in the boundaries of the nagi-nata in the experiment of the last rows of Fig. 4.

6 CONCLUSIONS

We have proposed a variational model for the joint estimation of the scene flow and its associated motion occlusions. Our work stems from the classical scene flow model presented in (Wedel et al., 2011) and incorporates a characterization of the occlusion areas as well as new data terms. The estimation of the occlusion map is useful to select a different set of data

Table 1: Results in MPI-Sintel training set for the optical flow (u,v) and for the disparity change δd. The first and second set of results correspond, respectively, to the *Final* and *Clean* frames. *EPE* means endpoint error over the complete frames. *EPE-M* shows the endpoint error over regions that remain visible in adjacent frames. *EPE-U* shows the endpoint error over regions that are visible only in one of the two adjacent frames. Notice that the ground truth δd is not provided in the database. We have set $\delta d(x,y) = \delta d_{t+1}(x+u, y+v) - d(x,y)$ using the (u,v,d_t,d_{t+1}) ground truth values. Using that information we have obtained the *EPE*-δd for all the image.

	EPE	EPE-M	EPE-U	EPE-δd
Final				
Classic Wedel	9.1461	7.7189	17.7888	1.1234
Our Wedel	7.6287	5.3934	19.8561	0.8121
Our Proposal	7.5095	5.2406	18.9948	0.7997
Clean				
Classic Wedel	8.6722	7.2324	17.2608	1.03522
Our Wedel	4.5097	2.2905	15.5042	0.5634
Our Proposal	4.3041	2.1558	15.1603	0.5521

Table 2: Results in KITTI 2015 training dataset for the optical flow (u,v) and for the disparity change δd. Out-noc (resp. Out-all) refers to the percentage of pixels where the estimated optical flow presents an error above 3 pixels in non-occluded areas (resp. all pixels). Out-δd refers to the percentage of pixels where the estimated disparity change presents an error above 3 pixels in the pixels where the disparity is available.

	Out-noc	Out-all	Out-δd
Classic Wedel	45.8745	55.4356	42.8971
Our Wedel	24.4237	33.2209	31.8971
Our Proposal	23.5233	32.8576	30.7532

terms for the occluded pixels, i.e., data terms that depend on the views where these pixels might be visible. We also have extended the optimization method for optical flow problems presented in (Palomares et al., 2016) to the scene flow case. Experimental results show, both quantitative and qualitatively, the benefits of the proposed energy functional and the minimization strategy. As future work we plan to use regularization and data terms that better preserve the image boundaries and that are more robust to illumination changes.

ACKNOWLEDGEMENTS

The authors acknowledge partial support by TIN2015-70410-C2-1-R (MINECO/FEDER, UE) and by GRC reference 2014 SGR 1301, Generalitat de Catalunya.

REFERENCES

Ayvaci, A., Raptis, M., and Soatto, S. (2012). Sparse occlusion detection with optical flow. International Journal of Computer Vision, 97(3):322–338.

Ballester, C., Garrido, L., Lazcano, V., and Caselles, V. (2012). A tv-l1 optical flow method with occlusion detection. In Pinz, A., Pock, T., Bischof, H., and Leberl, F., editors, DAGM/OAGM Symposium, volume 7476 of Lecture Notes in Computer Science, pages 31–40. Springer.

Basha, T., Moses, Y., and Kiryati, N. (2013). Multi-view scene flow estimation: A view centered variational approach. International Journal of Computer Vision, 101(1):6–21.

Brox, T., Bruhn, A., Papenberg, N., and Weickert, J. (2004). High accuracy optical flow estimation based on a theory for warping. In European Conference on Computer Vision (ECCV), volume 3024 of Lecture Notes in Computer Science, pages 25–36. Springer.

Butler, D. J., Wulff, J., Stanley, G. B., and Black, M. J. (2012). A naturalistic open source movie for optical flow evaluation. In European Conference on Computer Vision, pages 611–625.

Cech, J., Sanchez-Riera, J., and Horaud, R. P. (2011). Scene flow estimation by growing correspondence seeds. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3129–3136.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In Conference on Computer Vision and Pattern Recognition (CVPR).

Huguet, F. and Devernay, F. (2007). A variational method for scene flow estimation from stereo sequences. In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pages 1–7.

Ince, S. and Konrad, J. (2008). Occlusion-aware optical flow estimation. IEEE Transactions on Image Processing, 17(8):1443–1451.

Jaimez, M., Souiai, M., Stueckler, J., Gonzalez-Jimenez, J., and Cremers, D. (2015). Motion cooperation: Smooth piece-wise rigid scene flow from rgb-d images. In Proc. of the Int. Conference on 3D Vision (3DV). ¡a href=”https://youtu.be/qjPsKb-˙kvE”target=”˙blank”¿[video]¡/a¿.

Menze, M. and Geiger, A. (2015). Object scene flow for autonomous vehicles. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Palomares, R. P., Meinhardt-Llopis, E., Ballester, C., and Haro, G. (2016). Faldoi: A new minimization strategy for large displacement variational optical flow. Journal of Mathematical Imaging and Vision, pages 1–20.

Pons, J. P., Keriven, R., and Faugeras, O. (2007). Multi-view stereo reconstruction and scene flow estimation

with a global image-based matching score. International Journal on Computer Vision, 72(2):179–193.

Quiroga, J., Brox, T., Devernay, F., and Crowley, J. (2014). Dense semi-rigid scene flow estimation from rgbd images. In ECCV 2014, pages 567–582.

Sand, P. and Teller, S. (2008). Particle video: Long-range motion estimation using point trajectories. International Journal of Computer Vision, 80(1):72–91.

Sun, D., Sudderth, E. B., and Pfister, H. (2015). Layered rgbd scene flow estimation. In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, pages 548–556.

Vedula, S. and et al. (1999). Three-dimensional scene flow. In Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on. IEEE, volume 2, pages 722–729.

Vedula, S., Rander, P., Collins, R., and Kanade, T. (2005). Three-dimensional scene flow. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(3):475–480.

Vogel, C., Schindler, K., and Roth, S. (2011). 3d scene flow estimation with a rigid motion prior. In Computer Vision (ICCV), 2011 IEEE International Conference on, pages 1291–1298.

Vogel, C., Schindler, K., and Roth, S. (2015). 3d scene flow estimation with a piecewise rigid scene model. International Journal of Computer Vision, 115(1):1–28.

Wang, Y., Zhang, J., Liu, Z., Wu, Q., Chou, P. A., Zhang, Z., and Jia, Y. (2015). Handling occlusion and large displacement through improved rgb-d scene flow estimation. IEEE Transactions on Circuits and Systems for Video Technology, 26(7):1265–1278.

Wedel, A., Brox, T., Vaudrey, T., Rabe, C., Franke, U., and Cremers, D. (2011). Stereoscopic scene flow computation for 3d motion understanding. International Journal of Computer Vision, 95(1):29–51.

Weinzaepfel, P., Revaud, J., Harchaoui, Z., and Schmid, C. (2013). DeepFlow : Large displacement optical flow with deep matching. International Conference on Computer Vision.

Young, D. M. (1971). Iterative solution of large linear systems. Computer science and applied mathematics. Academic Press, Orlando.

Zanfir, A. and Sminchisescu, C. (2015). Large displacement 3d scene flow with occlusion reasoning. In Proceedings of the IEEE International Conference on Computer Vision, pages 4417–4425.