Dense Scene Flow Based on Depth and Multi-channel Bilateral Filter

Xiaowei Zhang, Dapeng Chen, Zejian Yuan, and Nanning Zheng

Institute of AI & Robotics, Xi'an Jiaotong University
zxwxjtu2010@gmail.com, chendapeng1988@stu.xjtu.edu.cn,
{zjyuan,nnzheng}@mail.xjtu.edu.cn

Abstract. There is close relationship between depth information and scene flow. However, it's not fully utilized in most of scene flow estimators. In this paper, we propose a method to estimate scene flow with monocular appearance images and corresponding depth images. We combine a global energy optimization and a bilateral filter into a two-step framework. Occluded pixels are detected by the consistency of appearance and depth, and the corresponding data errors are excluded from the energy function. The appearance and depth information are also utilized in anisotropic regularization to suppress over-smoothing. The multi-channel bilateral filter is introduced to correct scene flow with various information in non-local areas. The proposed approach is tested on Middlebury dataset and the sequences captured by KINECT. Experiment results show that it can estimate dense and accurate scene flow in challenging environments and keep the discontinuity around motion boundaries.

1 Introduction

Scene flow, which is also known as 3D motion field, is an important characteristic of dynamic scenes. It's quite useful for the applications such as object detection, segmentation, 3D reconstruction, tracking, virtual reality and so on.

Many approaches [1] [2] [3]are proposed to estimate scene flow on a stereo or multi-view camera system, in which the structure and scene flow are estimated simultaneously. It is quite difficult because observations on image planes are highly ambiguous, especially for the areas with weak textures.

With the development of sensor technologies, it's possible to get dense and accurate depth data in various ways, such as time-of-flight cameras, structured light cameras and the combination of these cameras with color camera systems [5]. So it's no longer necessary to estimate structure as before. Besides, structure information can be used to improve the accuracy of scene flow. So in this paper, we focus on estimating more precise scene flow with appearance images and corresponding depth images.

We combine the global energy optimization and a multi-channel bilateral filter into a two-step updating framework. The occluded pixels are detected according to the consistency of appearance and depth, and low weights are assigned to the

K.M. Lee et al. (Eds.): ACCV 2012, Part III, LNCS 7726, pp. 140–151, 2013.

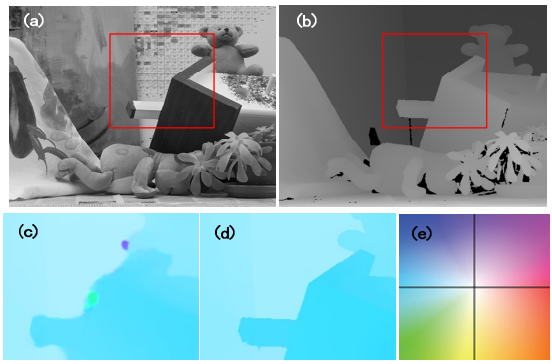

Fig. 1. Comparison of optical flow on Teddy in Middlebury stereo dataset. (a) The reference appearance image. (b) The ground truth of depth. (c) Optical flow of pixels in red box estimated by Brox[4]. (d) Optical flow of pixels in red box estimated by proposed method. (e)Color map of optical flow.

corresponding data term of energy function. To suppress over-smoothing in the scene flow, we propose an anisotropic regularization for the energy function. It imposes weak smoothness constraint along the direction in which the appearance and depth of pixels change intensely. The result obtained from global energy optimization is passed to the multi-channel bilateral filter, which can reject outliers and improve the accuracy of scene flow, especially around motion boundaries. A sample result is shown in Figure 1.

1.1 Related Works

There are some efficient local methods that provide dense or semi-dense scene flow. [6] presents a simple seed growing algorithm for scene flow estimation on a stereo setup. The disparity and scene flow of seeds are estimated first and then propagated to their neighborhood. [7] make use of RGB-D cameras and model the moving points in 3D space using particle filter. It supports multiple hypotheses and avoid over-smoothing the motion field. These local methods are efficient and have the ability to capture large motion, but the scene flow is not accurate and robust enough due to the lack of constraints among the adjacent pixels.

The most common methods to estimate scene flow is based on the optimization of global energy function, which includes brightness constancy matching and some regularization. The algorithms proposed in [4] [1] estimate structure and scene flow simultaneously in a stereo camera system. Basha et al. [3] showed that estimation could be improved by formulating the problem as a point cloud in 3D space, rather than the commonly used 2D parameterizations, where smoothness constraints are less applicable. Their method can be applied to any setup with multiple cameras. All these methods based on global energy optimization provide dense results even if the texture is quite weak.

However, the unreasonable regularization in the energy function result in over-smoothing in scene flow, which is an important error source of these methods. Many approaches are proposed to solve the similar problem in optical flow estimation with the relationship between appearance and flow. [8] and [9] applied anisotropic regularization in optical flow estimation, then it is introduced in scene flow estimation[10] by applying smoothness constraints only within segments in the image plane. [11] and [12] apply bilateral filter in the optical flow estimation, and the filter is proved to be the key factor in the accuracy improvement. However, to the best of our knowledge few methods make use of the relationship between structure and scene flow in a bilateral filter, which is usually more reliable than the relationship between appearance and scene flow. That's one of the most important difference in our approach.

In addition to the image plane and 3D point cloud, some other expressions of scene flow are proposed. [13] and [14] make use of 3D formulations based on meshes, which are less generally applicable than point clouds. [15] express the scene as 3D voxel model, but the accuracy is limited due to the discretization. Besides, [16] [17] and [18] have made some attempts to incorporate depth sensor into scene flow estimation.

2 The Method

2.1 Framework

Given color images I_t^0, I_{t+1}^0, and their corresponding disparity maps D_t, D_{t+1}, the scene flow to be estimated is referred as U_t, V_t and W_t. Here U_t and V_t are the motion along X and Y axis of camera coordinate system, and W_t is the change of disparity. For a field at time t, the subscript t is ignored in some places for simplicity. $p = (x, y)$ refers to a lattice of a 2D field F, and the corresponding value of the lattice p can be accessed by F_p.

To estimate the initial scene flow caused by ego-motion of the camera, the transformation of the camera is computed from match pairs of SURF features [19]. Then the energy function is build and solved with variational method, and the scene flow is processed with a multi-channel bilateral filter. The filtered scene flow is used as the initial scene flow for the next iteration. The steps are carried out iteratively until the scene flow is converged.

2.2 Energy Function

The consistency of brightness and depth is the basic assumption in scene flow estimation. However, brightness constancy assumption may be violated due to random noise and the change of lighting or exposure parameters. Some methods based on structure-texture decomposition are proposed to deal with brightness constancy violation [20], but they are time-consuming. Here we use weighted multi-channel images I in data energy, including a brightness channel, two chromatic aberration channels and two gradient channels. The weight of each channel

is determined by the average deviation of respective channel at time t and $t+1$. The channels that have large errors are assigned low weights. Its simple but effective in practical application. The energy of data deviation is expressed as follows:

$$\begin{aligned} E_{data}(U,V,W) = &\sum_{p\in\Omega} C_1(p)\varphi((I_{t+1}(p+\Delta p) - I(p))^2) \\ &+ \alpha_2 \sum_{p\in\Omega} (C_2(p)\varphi(D_{t+1}(p+\Delta p) - D(p) - W(p))^2) \end{aligned} \tag{1}$$

Here $\Delta p = (U(p), V(p))^T$. C_1 and C_2 are introduced to depress the influence of occluded pixels, which will be discussed below. $\varphi(\mathrm{x}^2) = \sqrt{\mathrm{x}^2+\epsilon^2}$ is a robust function that can reduce the influence of outliers, where $\epsilon = 0.001$.

An anisotropic regularization(smoothness constraint) is integrated into energy function to deal with ambiguities and noise of data. The regularization item is defined as:

$$\begin{aligned} E_{smooth}(U,V,W) = &\alpha_3 \sum_{p\in\Omega} \phi(\delta_p(\nabla U^T R \nabla U + \nabla V^T R \nabla V)) \\ &+ \alpha_4 \sum_{p\in\Omega} \phi(\delta_p(\nabla W^T R \nabla W)) \end{aligned} \tag{2}$$

Here $\delta_p(F)$ is the function that samples the value of a field F at p, and R is the anisotropic diffusion tensor operator which will be discussed below. $\nabla = (\partial_x, \partial_y)$, α_3 and α_4 are constant values that determine the weight of regularization. Robust function $\phi(\mathrm{x}^2) = \sqrt{\mathrm{x}^2+\epsilon^2}$ is introduced to keep scene flow discontinuous around motion boundaries. The total energy $E(U,V,W)$ is defined as the integration of data energy and regularization: $E(U,V,W) = E_{data}(U,V,W) + E_{smooth}(U,V,W)$

The estimation of scene flow is equivalent to minimization of energy function, so the scene flow can be obtained by solving Euler-Lagrange equations. If U and V are small enough, $E(U,V,W)$ can be linearized to simplify the optimization. Let $I_x = \partial I_{t+1}/\partial x$, $I_y = \partial I_{t+1}/\partial y$, $I_z = I_{t+1} - I_t$, D_x,D_y and D_z are similar.

Let $\varepsilon_I = I_x U + I_y V + I_z$, $\varepsilon_D = D_x U + D_y V + D_z - W$

The Euler-Lagrange equations can be expressed as:

$$\begin{aligned} \frac{\partial E}{\partial U} = &C_1 * \Psi'(\varepsilon_I^2) * I_x * \varepsilon_I + \alpha_2 C_2 * \Psi'(\varepsilon_D^2) * D_x * \varepsilon_D \\ &-\alpha_3 div(\Phi'(\nabla U^T R \nabla U + \nabla V^T R \nabla V) * (R \nabla U)) = 0 \end{aligned} \tag{3}$$

$$\begin{aligned} \frac{\partial E}{\partial V} = &C_1 * \Psi'(\varepsilon_I^2) * I_y * \varepsilon_I + \alpha_2 C_2 * \Psi'(\varepsilon_D^2) * D_y * \varepsilon_D \\ &-\alpha_3 div(\Phi'(\nabla U^T R \nabla U + \nabla V^T R \nabla V) * (R \nabla V)) = 0 \end{aligned} \tag{4}$$

$$\begin{aligned}\frac{\partial E}{\partial W} =& -\alpha_2 C_2 * \Psi'(\varepsilon_D^2) * D_x * \varepsilon_D \\ & -\alpha_4 div(\Phi'(\nabla W^T R \nabla W) * (R \nabla W)) = 0\end{aligned} \tag{5}$$

Here the operator * means that for every element in the left matrix, perform multiplication with the corresponding element in the right matrix. $\Psi'(F^2)$ is defined as $\Psi'(F^2)|_p = 0.5/\sqrt{F(p)^2 + \epsilon^2}$, $\Phi'(F^2)$ is the same as $\Psi'(F^2)$ in this paper. The Euler-Lagrange equations can be solved by SOR method.

To deal with large scene flow, we adopt coarse-to-fine image warping in the method [21]. Although it doesn't work well for small structures with large motion [4], this case can be avoided by improving the frame rate of camera.

In this paper the energy function is optimized with variational method, which is based on the assumption that the appearance and depth change smoothly, so the image should be smoothed before scene flow estimation. However, It will result in the leak of scene flow around motion boundaries, which are possibly depth boundaries too. For the pixels around depth boundaries with much larger depth than nearby pixels, their scene flow have no relationship with the pixels on the boundaries and foreground. While for the pixels around depth boundaries with much smaller depth than nearby pixels, their motion has close relationship with boundaries. Their appearance and depth should be smoothed with all nearby pixels, including those with similar or much larger depth, so that their scene flow can be estimated with variational method. In our approach, when the appearance or depth are smoothed only the pixels with similar or much larger depth is considered. Let F be the field to be smoothed, $N(p)$ be the smooth window of p, than the smoothed value F^s is defined as:

$$F^s(p) = \sum_{q \in N(p)} w_{p,q} F(p) \; D(p) = \sum_{q \in N(p)} w_{p,q} D(p)$$

Here $G(x, \sigma)$ is a Gaussian function with mean value 0 and variance σ,

$$w_{p,q} = \begin{cases} G(|p-q|, \sigma_s), & if\, D_p \le D_q \\ G(|p-q|, \sigma_s)\left(0.5 - \tan^{-1}(\kappa \cdot (D_p - D_q - \varepsilon_s))/\pi\right), & if\, D_p > D_q \end{cases}$$

2.3 Occlusion Handling

There are mainly two kinds of occlusion: the first one is motion occlusion, which is generated due to object motion. The second one is due to mismatch, which happens under various conditions such as change of color, object appearing and disappearing, shadow, etc. We detect both kinds of occlusions by checking the consistency of appearance and depth at different time. For the sake of robustness, we define C_I and C_D for I and D respectively:

$$C_I(p) = 0.5 - \tan^{-1}(\kappa_I(I_{t+1}(p + \Delta p) - I(p))^2 - \kappa_I \epsilon_I)/\pi$$
$$C_D(p) = 0.5 - \tan^{-1}(\kappa_D(D_{t+1}(p + \Delta p) - D(p) - W(p))^2 - \kappa_D \epsilon_D)/\pi$$

The matrix $C_1(p)$ and $C_2(p)$ in Equation (1) is defined as :
$C_1(p) = C_2(p) = max\{C_I(p)C_D(p), \epsilon_{min}\}$

The values in C_1 and C_2 rank between 0 and 1, and the pixels with large difference in appearance or depth are assigned small values, indicating that the regularization in energy function dominates the motion of these pixels.

2.4 Anisotropic Regularization

Regularization is usually integrated in the energy function to deal with ambiguities and noise of data. However, it may result in severe over-smoothing in the scene flow around motion boundaries due to unreasonable constrains. The adjacent pixels whose motion are quite different are likely to have different appearance and depth. The latter relationship between scene flow and depth is much more reliable in most cases. We propose an anisotropic regularization in the energy function to suppress over-smoothing around motion boundaries. It imposes weak smooth constraint along the directions in which the appearance or depth of the pixels changes intensely. For any pixel (x,y), the diffusion tensor operator is a 2*2 matrix defined as follows:

$$\begin{aligned} R_{11}(x,y) = &(w_I I_y(x,y)^2 + (1-w_I)D_y(x,y)^2 + \gamma^2) \\ &/(w_I I_x(x,y)^2 + w_I I_y(x,y)^2 + (1-w_I)D_x(x,y)^2 \\ &+ (1-w_I)D_y(x,y)^2 + 2\gamma^2) \end{aligned} \tag{6}$$

$$\begin{aligned} R_{22}(x,y) = &(w_I I_x(x,y)^2 + (1-w_I)D_x(x,y)^2 + \gamma^2) \\ &/(w_I I_x(x,y)^2 + w_I I_y(x,y)^2 + (1-w_I)D_x(x,y)^2 \\ &+ (1-w_I)D_y(x,y)^2 + 2\gamma^2) \end{aligned} \tag{7}$$

$$R_{12}(x,y) = R_{21}(x,y) = 0 \tag{8}$$

Here constant value w_I and $1-w_I$ determine the weights of constraints estimated by appearance and depth respectively. w_I is small than 0.5 since the constraints got from depth are more reliable. γ controls the degree of anisotropic, and for any pixel p there is at least one pixel that has close relationship with p in scene flow.

2.5 Multi-channel Bilateral Filter

Although we get dense and smooth scene flow by optimizing energy function, and the over-smoothing is suppressed to a great extent due to the anisotropic regularization, the result is still not satisfactory. Firstly, the motion of pixels around motion boundaries are not sharp as expected and varies a lot with parameters in the regularization. If the parameters are not proper, outliers may appear and have negative influence to the results. Secondly, in the energy function only the pair-wise smoothness of adjacent pixels is considered, so it can't utilize the information of the pixels that are not directly connected.

To cover the shortage, a multi-channel bilateral filter is introduced to correct scene flow with various information of non-local pixels. It can be considered as an expansion of anisotropic regularization, but it's much more powerful. Here we express a Gaussian function with mean value 0 and variance σ as $G(x, \sigma)$. Let

$G_I(p, p') = G(I(p) - I(p'), \sigma_I)$
$G_D(p, p') = G(D(p) - D(p'), \sigma_D)$
$G_d(p, p') = G(|p - p'|, \sigma_d)$
$G_O(p, p') = 1 - \mu + \mu * G(\sqrt{(U(p) - U(p'))^2 + (V(p) - V(p'))^2}, \sigma_O)$

G_I, and G_D are the weights determined by appearance and depth. G_d is the weight corresponding to the distance on the image plane, and near pixels have large weights. G_O is the weight got from optical flow, and it has limited influence due to the minimum value $1 - \mu$.

Let $M(p) = C_1(p)G(\sqrt{|\Delta U(p)|^2 + |\Delta V(p)|^2}, \sigma_M)$. Pixels with large $M(p)$ is likely to be occluded or mismatched. For every pixel, the weight W_{sum} is defined as:

$$W_{sum} = \sum_{p' \in N_f(p)} \{G_I(p, p')G_D(p, p')G_d(p, p')G_O(p, p')M(p')\} \tag{9}$$

Here $N_f(p)$ is the neighborhood of p.

Given a field F, the filtered value $\hat{F}(p)$ can be computed by:

$$\hat{F}(p) = \sum_{p' \in N_f(p)} \{F(p')G_I(p, p')G_D(p, p')G_d(p, p')G_O(p, p')M(p')\}/W_{sum} \tag{10}$$

Before filtering, all the pixels with large M are marked 0, and the rest pixels are marked 1. The filtering is divided into two stages. In the first stage only the points marked 0 is filtered. The area marked 0 may be quite large. If the size of filter window is too small, not all the pixels marked 0 will be filtered. If large enough to cover the area, the accuracy will be reduced seriously. In this paper we filter the scene flow iteratively with a fixed window size, until all the pixels that fulfill certain requirement are filtered. In every iteration, only the pixels that are marked 0 and have large W_{sum} will be filtered. If a pixel is successfully filtered, mark it to 1 and update the corresponding M, so that it can be used in the following procedure. In the second stage, a similar bilateral filtering with a small window size is carried out over all the scene flow.

3 Experiment

3.1 Middlebury Dataset

In order to perform quantitative evaluation, we test the method on Middlebury dataset. The images are captured by a set of cameras which are parallel and equally spaced along the X axis at the same time. The motion along Y and Z axis is always zero, and the ground truth of motion along X axis can be obtained from corresponding disparity, which is available in the Middlebury dataset.

Our approach is compared with four other methods, and the methods that only provide sparse scene flow [6] [7] are not considered. In the experiments Brox2010[4] + depth and Sun2010[12] + depth, we estimate optical flow with only monocular appearance images, and then compute scene flow directly using the disparity. In the experiments Huguet2007[1] and Basha2010[3] nothing is changed. The input disparity for Brox2010[4] + depth, Sun2010[12] + depth and our method is provided by [22]. For Huguet2007[1] and Basha2010[3], the disparity is estimated simultaneously in the method. In the experiment all the methods use their default parameters.

There are two kinds of criteria to evaluating scene flow. The first one projects 3D flow to 2D flow on image plane and computes the statistical deviation from 2D ground truth. The accuracy of optical flow and change of depth are evaluated respectively. The other criteria directly compute the statistical deviation of estimated flow from the ground truth in 3D space. For the systems that get depth information by stereo cameras or structured light cameras, the accuracy of depth decreases with the increase of distance, so the first evaluation criteria is less sensitive to the increasing noise of depth. Besides, human is much more sensitive to the near motion, which is more consistent with the first criteria. While the second evaluation criteria is more suitable for applications like 3D construction, which emphasize the accuracy of result in global coordinate system. In this paper we choose the first evaluation criteria. As described in [3], we evaluate the accuracy of optical flow with NRMS error. Since the ground truth of motion along Z axis is zero, we use RMS error to evaluate accuracy of w. AAE is used to evaluate direction of flow.

Some results are shown in Table 1. We can see that our method outperforms the rest methods in $NRMS_O$ in most cases, but the improvement is not as great as expected. An important reason is that the proposed method imposes strong constraints in the areas with weak textures and smooth disparities, and the motion in these areas will spread to a much larger area than other methods. If these areas are under the influence of brightness variations and weak moving shadows, the errors of scene flow are likely to be greater than the other methods. So although the result of our method is more accurate around motion boundaries, it doesn't outperform significantly in $NRMS_O$. The AAE of our method is not the best for similar reason. Besides, Table 1 shows that none of the compared methods outperform the others obviously in AAE, and we argue that AAE is not stable enough as an evaluation criterion. $NRMS_W$ are not compared here because the accuracy of motion in depth depend on the disparity images to a great extent, which is not the emphasis of this paper. Besides, if the disparity image and optical flow is accurate enough, it's easy to estimate accurate motion in depth.

Figure 2 shows some of the scene flow. In the magnified image we can see that there are over-smoothing in the flow estimated by Brox2010[4]+depth, Huguet2007[1] and Basha2010[3]. The result of Sun2010[12]+depth keeps the discontinuity of scene flow, but the motion of foreground leaks into some areas with weak texture, and the boundaries of flow is not smooth as is expected. The

Table 1. The evaluated errors of compared methods

	Methods	$NRMS_O$	AAE
Cones	Basha2010	3.07	**0.39**
	Brox2010+depth	1.32	0.60
	Sun2010+depth	1.21	0.50
	Huguet2007	5.97	0.69
	ours	**1.04**	0.69
Teddy	Basha2010	2.85	1.01
	Brox2010+depth	1.32	0.42
	Sun2010+depth	0.78	**0.18**
	Huguet2007	6.21	0.51
	ours	**0.73**	0.66
Venus	Basha2010	1,98	1.58
	Brox2010+depth	0.38	**0.78**
	Sun2010+depth	0.20	0.88
	Huguet2007	3.70	0.98
	ours	**0.15**	1.15

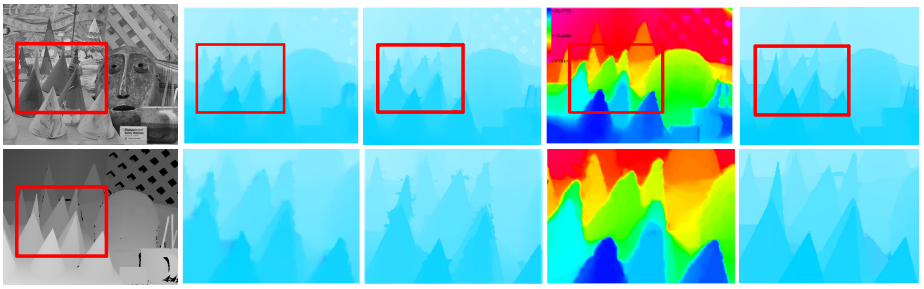

Fig. 2. Comparison of optical flow on an image in Middlebury dataset. The ground truth of optical flow is proportional to the depth. **Col 1**: Reference image and depth map given by Middlebury stereo dataset. **Col 2**:Brox2010 [4]. **Col 3**:Sun2010[12]. **Col 4**:Huguet2007[1]. Although the optical flow is colored in the way described in [1], it clearly shows overs-smoothing around boundaries. **Col 5**:our method. The images in top row are the optical flow of the whole image, and the images in bottom row are the optical flow in the red box above.

scene flow estimated by our approach is more accurate, especially around motion boundaries.

Qualitative evaluation is performed on appearance and depth sequence captured by KINECT. The KINECT provides calibrated appearance and depth sequence at a fps of 30Hz, and their resolutions are both 640*480. The range of depth measurement is from 0.5m to 5.5m, and the precision decreases with the increase of distance. The appearance image is influenced by noise, lossy compression, change of light and etc.

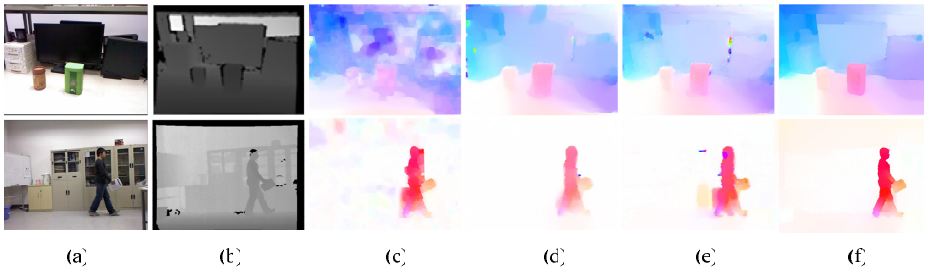

Fig. 3. Comparison of optical flow on our dataset. **(a)**: Color images. **(b)**: Depth maps obtained by KINECT. **(c)**: Sun[12]. **(d)**: Brox[4]. **(e)**: Our method without anisotropic regularization and bilateral filter. **(f)**: our method.

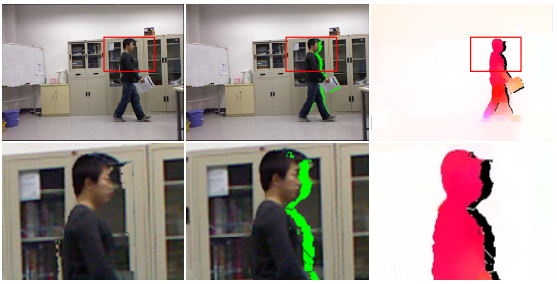

Fig. 4. Left column: warped image according to the optical flow by Sun2010[12]+depth. Middle column: warped image according to the optical flow by our approach. Occluded pixels are marked green. Right column: optical flow estimated by our method, in which the occluded pixels are marked black. The images in the bottom row show the close-up of the corresponding areas in red box above.

The top row of Figure 3 shows the experiment in which a static scene is viewed in different positions. There are many areas with weak texture, such as the desks and the wall. The bottom row of Figure 3 shows the experiment in which a man is walking in the room. The camera is stationary relative to the room. The man is a non-rigid object, whose optical flow is smooth but not constant. Unexpectedly, Sun2010[12]+depth doesn't work well with the default parameters. The motion of foreground leaks into many areas with weak texture, and many artificial boundaries of flow appears in the areas which have smooth motion. That's because Sun2010[12] is an image-driven approach, and the scene flow will spread widely in the areas with weak texture. While in the areas with rich texture and non-constant motion, artificial boundaries is likely to appear in the scene flow due to the bilateral filter.

From the results in Figure 3 we can see that there are serious over-smoothing in the flow estimated by [4]+depth and our method in which depth isn't considered in regularization and bilateral filter. Outliers also appear around motion boundaries. Our result is much more precise especially around motion

boundaries, and no outliers appear in the scene flow. It is clear that the utilization of depth information in regularization and bilateral filter improves the precision of scene flow significantly.

It should be noticed that the error of scene flow is large in areas with weak texture, such as the wall and the floor. An important reason is that the color image captured by KINECT doesn't meet the brightness constancy assumption exactly, and the light shadows or the change of lighting and camera parameters may cause false motion in these areas. For the methods using isotropic regularization, the scene flow in these areas is restrained by the boundaries of foreground. Although the result may be similar to the ground truth, it's just coincidence and doesn't mean these methods perform better than ours, because there is no relationship between the motion of foreground boundaries and the motion of background. Indeed, it's ambiguous that how to define the scene flow of those pixels which are located in areas lack of texture and influenced by false motion of appearance. In this sense, the ground truth of scene flow has more than one value and the result of our method is also "accurate".

Due to the lack of ground truth, we can't evaluate the precision of scene flow directly. Here we warp the color images to the plane of reference image, and then check the consistency of appearance. The occluded pixels which are detected by the approach are ignored. Figure shows that the scene flow estimated by our approach is more accurate around motion boundaries, and the occlusion detection is accurate enough for other use.

4 Conclusions

In this paper, we propose a robust method to estimate high-accurate scene flow based on appearance and depth information. This method combines global energy optimization and multi-channel bilateral filter into a two-step framework. The former component provides dense scene flow and suppress over-smoothing around motion boundaries. The latter component correct scene flow with information in non-local area, including appearance, depth, distance on image plane and flow. Experiment results show that with depth information considered in regularization and bilateral filter, our method successfully provides accurate scene flow in challenging environment.

Acknowledgement. This work was supported in part by the National Basic Research Program of China under Grant No. 2012CB316400, and the National Natural Science Foundation of China under Grant No. 91120006.

References

1. Huguet, F., Devernay, F.: A variational method for scene flow estimation from stereo sequences. In: ICCV, pp. 1–7. IEEE (2007)
2. Wedel, A., Brox, T., Vaudrey, T., Rabe, C., Franke, U., Cremers, D.: Stereoscopic Scene Flow Computation for 3D Motion Understanding. IJCV, 1–23 (2010)

3. Basha, T., Moses, Y., Kiryati, N.: Multi-view scene flow estimation: A view centered variational approach. In: CVPR, pp. 1506–1513. IEEE (2010)
4. Brox, T., Malik, J.: Large displacement optical flow: Descriptor matching in variational motion estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 500–513 (2011)
5. Zhu, J., Wang, L., Yang, R., Davis, J.: Fusion of time-of-flight depth and stereo for high accuracy depth maps. In: CVPR, pp. 1–8. IEEE (2008)
6. Jan, Sanchez-Riera, J., Horaud, R.: Scene flow estimation by growing correspondence seeds. In: CVPR, pp. 3129–3136. IEEE (2011)
7. Hadfield, S., Bowden, R.: Kinecting the dots: Particle based scene flow from depth sensors. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2290–2295. IEEE (2011)
8. Nagel, H., Enkelmann, W.: An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence, 565–593 (1986)
9. Zimmer, H., Bruhn, A., Weickert, J., Valgaerts, L., Salgado, A., Rosenhahn, B., Seidel, H.-P.: Complementary Optic Flow. In: Cremers, D., Boykov, Y., Blake, A., Schmidt, F.R. (eds.) EMMCVPR 2009. LNCS, vol. 5681, pp. 207–220. Springer, Heidelberg (2009)
10. Li, R., Sclaroff, S.: Multi-scale 3d scene flow from binocular stereo sequences. Computer Vision and Image Understanding 110, 75–90 (2008)
11. Xiao, J., Cheng, H., Sawhney, H.S., Rao, C., Isnardi, M.: Bilateral Filtering-Based Optical Flow Estimation with Occlusion Detection. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 211–224. Springer, Heidelberg (2006)
12. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: CVPR, pp. 2432–2439. IEEE (2010)
13. Courchay, J., Pons, J.-P., Monasse, P., Keriven, R.: Dense and Accurate Spatio-temporal Multi-view Stereovision. In: Zha, H., Taniguchi, R.-i., Maybank, S. (eds.) ACCV 2009, Part II. LNCS, vol. 5995, pp. 11–22. Springer, Heidelberg (2010)
14. Furukawa, Y., Ponce, J.: Dense 3d motion capture from synchronized video streams. Image and Geometry Processing for 3-D Cinematography, 193–211 (2010)
15. Vedula, S., Baker, S., Kanade, T.: Image-based spatio-temporal modeling and view interpolation of dynamic events. ACM Transactions on Graphics (TOG) 24, 240–261 (2005)
16. Spies, H., Jähne, B., Barron, J.: Range flow estimation. Computer Vision and Image Understanding 85, 209–231 (2002)
17. Lukins, T.C., Fisher, R.B.: Colour constrained 4d flow. In: Proceedings of the British Machine Vision Conference, pp. 340–348 (2005)
18. Schuchert, T., Aach, T., Scharr, H.: Range Flow for Varying Illumination. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 509–522. Springer, Heidelberg (2008)
19. Lepetit, V., Moreno-Noguer, F., Fua, P.: Epnp: An accurate o (n) solution to the pnp problem. International Journal of Computer Vision 81, 155–166 (2009)
20. Wedel, A., Pock, T., Zach, C., Bischof, H., Cremers, D.: An improved algorithm for tv-l 1 optical flow. Statistical and Geometrical Approaches to Visual Motion Analysis, 23–45 (2009)
21. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High Accuracy Optical Flow Estimation Based on a Theory for Warping. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 25–36. Springer, Heidelberg (2004)
22. Bleyer, M., Gelautz, M.: Graph-cut-based stereo matching using image segmentation with symmetrical treatment of occlusions. Signal Processing: Image Communication 22, 127–143 (2007)