Simultaneous Stereo Video Deblurring and Scene Flow Estimation

Liyuan Pan [1,2], Yuchao Dai[2], Miaomiao Liu[3,2] and Fatih Porikli[2]
[1] School of Automation, Northwestern Polytechnical University, Xi'an, China
[2] Research School of Engineering, Australian National University, Canberra, Australia
[3] Data61, CSIRO, Canberra, Australia
panliyuan@mail.nwpu.edu.cn, {yuchao.dai, miaomiao.liu, fatih.porikli}@anu.edu.au

Abstract

Videos for outdoor scene often show unpleasant blur effects due to the large relative motion between the camera and the dynamic objects and large depth variations. Existing works typically focus monocular video deblurring. In this paper, we propose a novel approach to deblurring from stereo videos. In particular, we exploit the piece-wise planar assumption about the scene and leverage the scene flow information to deblur the image. Unlike the existing approach [31] which used a pre-computed scene flow, we propose a single framework to jointly estimate the scene flow and deblur the image, where the motion cues from scene flow estimation and blur information could reinforce each other, and produce superior results than the conventional scene flow estimation or stereo deblurring methods. We evaluate our method extensively on two available datasets and achieve significant improvement in flow estimation and removing the blur effect over the state-of-the-art methods.

1. Introduction

Image deblurring aims at recovering latent clean images from a single or multiple images, which is a fundamental task in image processing and computer vision. Image blur could be caused by various reasons, for example, optical aberration [30], medium perturbation [18], temperature variation [23], defocus [33], and motion [4, 10, 17, 34, 42]. The blur not only reduces the quality of the image causing loss of important details, but also hampers further analysis. Image deblurring has been extensively studied and various methods have been proposed.

In this work, we focus on image blur caused by motion. Motion blur is widely encountered in real world applications such as autonomous driving [5, 7]. Camera and object motion blur effects become more apparent when the exposure time of the camera increases due to low-light conditions. It is common to model the blur effect using kernels [17, 21]. Under motion blur, the induced blur kernel would be in 2D [16] or 3D [31]. For a scenario where both camera motion and multiple moving objects exist, the blur

(a) Original blur images

(b) Kim and Lee CVPR 2015 [16]

(c) Sellent *et al*. ECCV 2016 [31]

(a) Ours

Figure 1. Stereo deblurring results on outdoor scenarios. (a) Two samples from the KITTI autonomous driving benchmark dataset. (b) Deblurring result of [16]. (c) Deblurring result of [31]. (d) Our deblurring result. Compared with both state-of-the-art monocular and stereo debluring methods, our method achieves the best performance especially for large motion regions in the scene. Best viewed in color on screen.

1063-6919/17 $31.00 © 2017 IEEE
DOI 10.1109/CVPR.2017.739

kernel is, in principle, defined for each pixel. Therefore, conventional blur removal methods, such as [2, 10, 26, 40] cannot be directly applied since they are restricted to a single or a fixed number of blur kernels, making them inferior in tackling general motion blur problems.

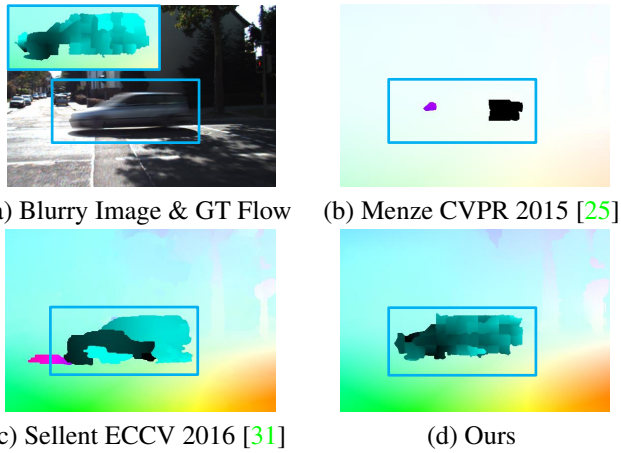
On another front, stereo-based depth and motion estimation have witnessed significant progress over the last decade thanks to the availability of large benchmark datasets such as Middlebury [29] and KITTI [7]. These benchmarks provide realistic scenarios with meaningful object classes and associated ground-truth annotations. The success of stereo-based motion estimation naturally prompts more advanced stereo based deblurring solutions, promising more accurate motion estimations to compensate for motion blurs. Very recently, Sellent *et al*. [31] proposed to exploit stereo information in aiding the challenging video deblurring task, where a piecewise rigid 3D scene flow representation is used to estimate motion blur kernels via local homographies. It makes a strong assumption that 3D scene flow can be reliably estimated, even under adverse conditions. While they reported favorable results on both synthetic and real data, all the experiments are confined to indoor scenarios.

The phenomenon around motion and blur can be viewed as a chicken-egg problem: More effective motion blur removal requires more accurate motion estimation. Yet, the accuracy of motion estimation highly depends on the quality of the images. We would like to argue that, scene flow estimation approaches that make use of color brightness constancy may be hindered by the blur images. In Fig. 2, we compare the scene flow estimation results of the state-of-the-art solutions on different blur images. It could be observed that the scene flow estimation performance deteriorates quickly w.r.t. the image blur.

Here, we aim to solve the above two problems simultaneously in a unified framework. Our motivation is that motion estimation and video deblur benefit from each other, *i.e*., better scene flow estimation will lead to a better deblurring result, and a cleaner image will lead to better flow estimation. We tackle a more general blur problem that is not only caused by camera motion but also by moving objects and depth variations in a dynamic scene. We define our problem as "generalized stereo deblur", where moving stereo cameras observe a dynamic scene with varying depths. We propose a new pipeline (see Fig. 4 for simultaneously estimating the 3D scene flow and deblurring images. Using our formulation, we attain significant improvement in numerous real challenging scenes as illustrated in Fig. 1.

The main contributions of our work are as follows:

- We propose a novel joint optimization framework to simultaneously estimate the scene flow and deblurred latent images for dynamic scenes. Our deblurring objective benefits from the improved scene flow estimates and the estimated scene structure. Similarly, the scene flow objective allows deriving more accurate pixel-wise spatially varying blur kernels.
- Based on the piece-wise planar assumption, we obtain a structured blur kernel model. More specifically, the optical flows for pixels in the same superpixel are constrained by a single homography (see Section.3.1).
- As our experiments demonstrate, our method can successfully handle complex real-world scenes depicting fast moving objects, camera motions, uncontrolled lighting conditions, and shadows.



(a) Blurry Image & GT Flow (b) Menze CVPR 2015 [25]

(c) Sellent ECCV 2016 [31] (d) Ours

Figure 2. Scene flow estimation results for an outdoor scenarios. (a) A blur frame from the KITTI autonomous driving benchmark datasets. (b) Estimated flow by [25]. (c) Estimated flow by [31]. (d) Our flow estimation result. Compared with both these state-of-the-art methods that rank the 1^{st} and 2^{nd} on the KITTI dataset, our method achieves the best performance especially for large motion region in the scene. Best viewed in color on screen.

2. Related Work

Blur removal is an ill-posed problem, thus certain assumptions or additional constraints are required to regularize the solution space. Numerous methods have been proposed to address this problem [16, 17, 31, 34], which can be categorized into two groups: monocular based approaches and binocular based approaches.

Monocular based approaches often assume that the captured scene is static and has a constant depth. Based on these assumptions, uniform or non-uniform blur kernels are estimated from a single image [10, 12, 14]. Hu *et al*. [14] proposed to jointly estimate the depth layering and remove non-uniform blur from a single blur image. While this unified framework is promising, user input for depth layers partition is required, and potential depth values should be known in advance. In practical settings, blur is spatially varying due to camera and object motion, which makes the kernel estimation a difficult problem.

Since blur parameters and the latent image are difficult to be estimated from a single image, the monocular based approaches are extended to video to remove blurs in dynamic scenes [32, 37]. To this end, Deng *et al*. [4] and He *et al*. [11] apply feature tracking of a single moving object to obtain 2D displacement-based blur kernels for deblurring. Matsushita *et al*. [24] and Cho *et al*. [3] proposed to exploit the existence of salient sharp frames in videos. Nevertheless, the method of Matsushita *et al*. [24] cannot remove blurs caused by moving objects. Moreover, the work of Cho [3] cannot handle fast moving objects which have distinct motions from those of backgrounds. Wulff and Black [38] proposed a layered model to estimate the different motions of both foreground and background layers. However, these motions are restricted to affine models, and it is difficult to be extended to multi-layer scenes due to the requirement of depth ordering of the layers.

Kim and Lee [15] proposed a method based on a local linear motion without segmentation. This method incorporates optical flow estimation to guide the blur kernel estimation and is able to deal with certain object motion blur. In [16], a new method is proposed to simultaneously estimate optical flow and tackle the case of general blur by minimization a single non-convex energy function. This method represents the state-of-the-art in video deblurring and is used for comparison in the experimental section.

As depth can significantly simplify the deblurring problem, the multi-view methods have been proposed to leverage on depth information. Building upon the work of Ezra and Nayar [28], Li *et al*. [22] extended the hybrid camera with an additional low-resolution video camera where two low-resolution cameras form a stereo pair and provide a low-resolution depth map. Tai *et al*. [35] used a hybrid camera system to compute a pixel-wise kernel with optical flow. Xu *et al*. [39] inferred depth from two blur images captured by a stereo camera and proposed a hierarchical estimation framework to remove motion blur caused by in-plane translation. Just recently, Sellent *et al*. [31] proposed a video deblurring technique based on stereo video, where 3D scene flow is estimated from blur images using a piecewise rigid 3D scene flow representation.

3. Formulation

Our goal is to handle the blurs in stereo videos caused by the motion of the camera, objects, and large depth variations in a scene. To this end, we formulate our problem as a joint estimation of scene flow and image deblurring for dynamic scenes. In particular, we rely on the assumptions that the scene can be approximated by a set of 3D planes [41] belonging to a finite number of objects[1] performing rigid motions [25]. Based on these assumptions, we define our structured blur kernel as well as the energy functions for deblurring in the following sections.

3.1. Blur Image Formation based on the Structured Pixel-wise Blur Kernel

Blur images are formed by the integration of light intensity emitted from the dynamic scene over the aperture time interval of the camera. This defines the image frame in the video sequence as

$$\mathbf{B}_m(\mathbf{x}) = \frac{1}{\tau}\int_{m-\frac{\tau}{2}}^{m+\frac{\tau}{2}} \mathbf{L}(m,\mathbf{x})\mathrm{d}m = \frac{1}{\tau}\int_{m-\frac{\tau}{2}}^{m+\frac{\tau}{2}} \mathbf{L}_m(\mathbf{x}+\mathbf{u}_m)\mathrm{d}m \tag{1}$$

where $\mathbf{B}_m$ is the blur frame, $\mathbf{L} \in [-T,T]\times\Omega$ is a continuous latent video sequence over a time interval $[-T,T]$, τ is the duty cycle, $\mathbf{u}_m$ is the optical flow at m. We denote $\mathbf{L}_m(\mathbf{x}) = \mathbf{L}(m,\mathbf{x})$. This leads to the discretized version of blur model in Eq. (1) as

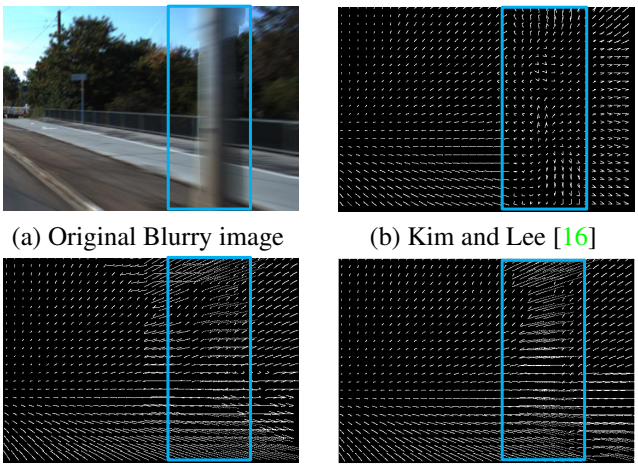
$$\mathbf{B}_m(\mathbf{x}) = \mathbf{A}_m^{\mathbf{x}}\mathbf{L}_m, \tag{2}$$

where $\mathbf{A}_m^{\mathbf{x}}$ is the blur kernel vector for the image at location $\mathbf{x}$. We obtain the blur kernel matrix $\mathbf{A}$ by stacking $\mathbf{A}^{\mathbf{x}}$. This leads to the blur model for the image as $\mathbf{B}_m = \mathbf{A}_m\mathbf{L}_m$. In order to handle multiple types of blurs, Kim *et al*. [15] approximated the pixel-wise blur kernel using bidirectional optical flows

$$k_{m,\mathbf{x}}(u,v) = \begin{cases} \dfrac{\delta(u\tilde{v}_{m+} - v\tilde{u}_{m+})}{\tau_m||\tilde{\mathbf{u}}_{m+}||}, & if\ u\in[0,\frac{\tau_m}{2}\tilde{u}_{m+}], v\in[0,\frac{\tau_m}{2}\tilde{v}_{m+}] \\ \dfrac{\delta(u\tilde{v}_{m-} - v\tilde{u}_{m-})}{\tau_m||\tilde{\mathbf{u}}_{m-}||}, & if\ u\in(0,\frac{\tau_m}{2}\tilde{u}_{m-}], v\in(0,\frac{\tau_m}{2}\tilde{v}_{m-}] \\ 0, & otherwise \end{cases} \tag{3}$$

where $k_{t,\mathbf{x}}$ is the blur kernel at $\mathbf{x}$, δ denotes the Kronecker delta, $\tilde{\mathbf{u}}_{m+} = (\tilde{u}_{m+},\tilde{v}_{m+})$ and $\tilde{\mathbf{u}}_{m-} = (\tilde{u}_{m-},\tilde{v}_{m-})$ are the bidirectional optical flows at frame m. In particular, $\mathbf{u}_{m+} = \mathbf{u}_{m\to m+1}$ and $\mathbf{u}_{m-} = \mathbf{u}_{m\to m-1}$. They jointly estimated the optical flow and the deblurred images. In our setup, the stereo video provides the depth information for each frame. Based on our piece-wise planar assumptions on the scene, optical flows for pixels lying on the same plane are constrained by a single homography. In particular, we represent the scene in terms of superpixels and finite number of objects with rigid motions. We denote $\mathcal{S}$ and $\mathcal{O}$ as the set of superpixels and moving objects, respectively. Each superpixel $i \in S$ is associated with a region $\mathcal{R}_i$ in the image with a plane variable $\mathbf{n}_{i,k} \in \mathbb{R}^3$ in 3D ($\mathbf{n}_{i,k}^T\mathbf{X} = 1$ for $\mathbf{X}\in\mathbb{R}^3$), where $k \in \{1,\cdots,|\mathcal{O}|\}$ denotes that superpixel i is associated with object k inheriting its corresponding motion parameters $\mathbf{o}_k = (\mathbf{R}_k,\mathbf{t}_k) \in \mathbb{SE}(3)$, where $\mathbf{R}_k \in \mathbb{R}^{3\times3}$ is the rotation matrix and $\mathbf{t}_k \in \mathbb{R}^3$ is the translation vector. Note that $(\mathbf{o}_k,\mathbf{n}_{i,k})$ encodes the scene

[1] The background can be regarded as a single 'object' due to the camera motion.



(a) Original Blurry image (b) Kim and Lee [16]

(c) Sellent *et al*. [31] (d) Ours

Figure 3. Blur kernel estimation on an outdoor scenario. (a) A blur frame from the KITTI autonomous driving benchmark datasets. (b) Blur kernel of [16]. (c) Blur kernel of [31]. (d) Our blur kernel. Compared with these monocular and stereo deblurring methods, our method achieves more accurate blur kernels.

flow information [25]. Given the parameters $(\mathbf{o}_k, \mathbf{n}_{i,k})$, we can obtain the homography defined for superpixel i as

$$\mathbf{H}_i = \mathbf{K}(\mathbf{R}_k - \mathbf{t}_k \mathbf{n}_{i,k}^T)\mathbf{K}^{-1} \tag{4}$$

where $\mathbf{K} \in \mathbb{R}^{3\times3}$ is the intrinsic matrix. The optical flow is then defined as

$$\mathbf{u}_{i,j} = \mathbf{H}_i \mathbf{x}_{i,j} - \mathbf{x}_{i,j} \tag{5}$$

where $\mathbf{x}_{i,j}$ is the coordinate of pixel j in superpixel i. This shows that the optical flows for pixels in a superpixel are constrained by the homography. Thus, it leads to a structured version of blur kernel defined in Eq. (3). In Fig. 3, we compare our blur kernel estimation with the Kim and Lee [16] and Sellent *et al*. [31]. Our kernels are more structural, which also leads to more accurate scene flow estimation.

3.2. Energy Minimization

We formulate the problem in a single framework as a discrete-continuous optimization problem to jointly estimate the scene flow and deblur the images. In particular, our energy minimization model is formulated as

$$\mathbf{E} = \underbrace{\sum_{i\in\mathcal{S}} \phi_i(\mathbf{n}_i, \mathbf{o}, \mathbf{L})}_{\text{data term}} + \underbrace{\sum_{i,j} \phi_{i,j}(\mathbf{n}_i, \mathbf{n}_j, \mathbf{o})}_{\substack{\text{scene flow}\\ \text{smoothness term}}} + \underbrace{\sum_{m} \phi_m(\mathbf{L})}_{\substack{\text{latent image}\\ \text{regularisation}}} \tag{6}$$

which consists of a data term, a smoothness term for scene flow, and a spatial regularization term for latent clean images. Our model is initially defined on three consecutive pairs of stereo video sequences. It can also allow the input with two pairs of frames. Details are provided in Section 5. The energy terms are discussed in the following sections.

In Section 4, we solve the optimization problem in an alternative manner to handle mixed discrete and continuous variables, thus allowing us to jointly estimate the scene flow and deblur the images.

3.3. Data Term

Our data term involves mixed discrete and continuous variables, and are of three different kinds. The first kind encodes the fact that the corresponding pixels across the six latent images should have similar appearance (brightness constancy). This lets us write the term as

$$\phi_i^1(\mathbf{n}_{i,k}, \mathbf{o}_k, \mathbf{L}) = \theta_1 |\mathbf{L}(\mathbf{x}) - \mathbf{L}^*(\mathbf{H}^*\mathbf{x})|_1, \tag{7}$$

where the superscript $* \in \{\mathbf{stereo}, \mathbf{flow}_{f,b}, \mathbf{cross}_{f,b}\}$ denotes the warping direction to other images and $(\cdot)_{f,b}$ denotes the forward and backward direction, respectively (see Fig. 4). We adopt the robust ℓ_1 norm to enforce its robustness against noise and occlusions.

Our second potential, similar to one term used in [25], is defined as

$$\phi_i^2(\mathbf{n}_{i,k}, \mathbf{o}_k) = \begin{cases} \theta_2 \rho_{\alpha_1}(||\mathbf{H}^*\mathbf{x} - \mathbf{x}^*||_2) & \text{if } \mathbf{x} \in \mathbf{\Pi}_\mathbf{x}, \\ 0 & \text{otherwise.} \end{cases}$$

where $\rho_\alpha(\cdot) = \min(|\cdot|, \alpha)$ denotes the truncated ℓ_1 penalty function. More specifically, it encodes the information that the warping of feature points $\mathbf{x} \in \mathbf{\Pi}_\mathbf{x}$ based on $\mathbf{H}^*$ should match its extracted correspondences in the target view. In particular, $\Pi_\mathbf{x}$ is obtained in a similar manner as [25].

The third data term, making use of the observed blur images, is defined as

$$\phi_i^3(\mathbf{n}_{i,k}, \mathbf{o}_k, \mathbf{L}) = \theta_3 \sum_m \sum_{\partial_*} \|\partial_* \mathbf{A}_m(\mathbf{n}_{i,k}, \mathbf{o}_k)\mathbf{L}_m - \partial_* \mathbf{B}_m\|_2^2$$

where ∂_* are the Toeplitz matrices corresponding to the horizontal and vertical derivative filters. This term encourages the intensity changes in the estimated blur images to be close to that of the observed blur images.

3.4. Smoothness Term for Scene Flow

Our energy model exploits a smoothness potential that involves the discrete and continuous variables. It is similar to the ones used in [25]. In particular, our smoothness term includes three different types. The first one is to encode the compatibility of two superpixels that share a common boundary by respecting the depth discontinuities. To this end, we define our potential function as

$$\phi_{i,j}^1(\mathbf{n}_i, \mathbf{n}_j) = \theta_4 \sum_{\mathbf{x}\in\mathcal{B}_{i,j}} \rho_{\alpha_2}(\omega_{i,j}(\mathbf{n}_i, \mathbf{n}_j, \mathbf{x})) \tag{8}$$

where $\omega(\mathbf{n}_i, \mathbf{x})$ is the disparity of pixel $\mathbf{x}$ in superpixel i in the reference disparity map, $\omega_{i,j}(\mathbf{n}_i, \mathbf{n}_j, \mathbf{x}) = \omega(\mathbf{n}_i, \mathbf{x}) -$

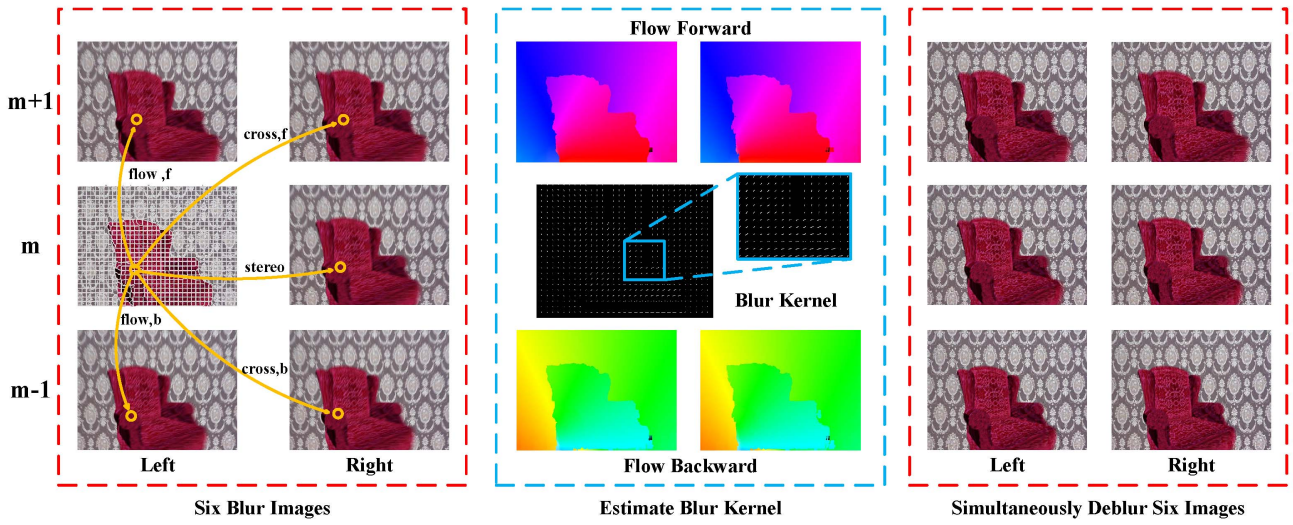


Figure 4. Illustration of our 'generalized stereo deblurring' method. We simultaneously compute four scene flows (in two directions and in two view), and deblur six images. In case the input contains only two images, we use the reflection of the flow forward as the flow backward in the deblurring part.

$\omega(\mathbf{n}_j, \mathbf{x})$ are the distance of disparity for pixel $\mathbf{x} \in \mathcal{B}_{i,j}$ on the boundary.

The second potential is to encourage the neighbor superpixels to orient in the same direction. It is expressed as

$$\phi^2_{i,j}(\mathbf{n}_i, \mathbf{n}_j) = \theta_5 \rho_{\alpha_3}\left(1 - \frac{|\mathbf{n}_i^T \mathbf{n}_j|}{\|\mathbf{n}_i\| \, \|\mathbf{n}_j\|}\right). \tag{9}$$

The third potential is to encode the fact that the motion boundaries are co-aligned with disparity discontinuities. This potential can be expressed as

$$\phi^3_{i,j}(\mathbf{n}_{i,k}, \mathbf{n}_{j,k'}) = \theta_6 \begin{cases} \exp\left(-\frac{\lambda}{|\mathcal{B}_{i,j}|} \sum\limits_{\mathbf{x} \in \mathcal{B}_{i,j}} \omega_{i,j}(\mathbf{n}_i, \mathbf{n}_j, \mathbf{x})^2 \frac{|\mathbf{n}_i^T \mathbf{n}_j|}{\|\mathbf{n}_i\| \|\mathbf{n}_j\|}\right) & \text{if } k \neq k', \\ 0 & \text{else.} \end{cases}$$

where $|\mathcal{B}_{i,j}|$ denotes the number of pixels shared along boundary between superpixels i and j.

3.5. Regularization Term for Latent Images

Spatial regularization has proven its importance in image deblurring [19, 20]. In our model, we use the total variation term to suppress the noise in the latent image while preserving edges, and penalize spatial fluctuations. Therefore, our potential takes the form

$$\phi_m = |\nabla \mathbf{L}_m|. \tag{10}$$

Note that the total variation is applied to each color channel.

4. Solution

The optimization of our energy function defined in Eq. (6), involving both discrete and continuous variables, is challenging to solve. Recall that our model involves two set of variables, namely scene flow variables and latent images. Fortunately, given one set of variables, we can solve the other efficiently. Therefore, we perform the optimization iteratively by the following steps,

- Fix latent image $\mathbf{L}$, solve scene flow by optimizing Eq. (11) (See Section 4.1).
- Fix scene flow parameters, $\mathbf{n}$ and $\mathbf{o}$, solve latent image by optimizing Eq. (12) (See Section 4.2).

In the following sections, we describe the details for each optimization step.

4.1. Scene flow estimation

We fix latent images, namely $\mathbf{L} = \tilde{\mathbf{L}}$, Eq. (6) reduces to

$$\min_{\mathbf{n},\mathbf{o}} \sum_{i \in \mathcal{S}} \phi_i^{1,2,3}(\mathbf{n}_i, \mathbf{o}, \tilde{\mathbf{L}}) + \sum_{i,j} \phi_{i,j}^{1,2,3}(\mathbf{n}_i, \mathbf{n}_j, \mathbf{o}). \tag{11}$$

which becomes a discrete-continuous CRF optimization problem. We use the sequential tree-reweighted message passing (TRW-S) method in [25] to find the solution.

4.2. Deblurring

Given the scene flow parameters, namely $\tilde{\mathbf{n}}$, and $\tilde{\mathbf{o}}$, the blur kernel matrix, $\mathbf{A}_m$ is derived based on Eq. (3), and Eq. (5). The objective function in Eq. (6) becomes convex with respect to $\mathbf{L}$ and is expressed as

$$\min_{\mathbf{L}} \sum_{i \in \mathcal{S}} \phi_i^1(\tilde{\mathbf{n}}_i, \tilde{\mathbf{o}}, \mathbf{L}) + \phi_i^3(\tilde{\mathbf{n}}_{i,k}, \tilde{\mathbf{o}}_k, \mathbf{L}) + \sum_m \phi_m(\mathbf{L}). \tag{12}$$

In order to obtain sharp image $\mathbf{L}$, we adopt the conventional convex optimization method [1] and derive the primal-dual updating scheme as follows

$$\begin{cases} \mathbf{p}_m^{r+1} = \dfrac{\mathbf{p}_m^r + \gamma \nabla \mathbf{L}_m^r}{\mathbf{max}(1, \mathbf{abs}(\mathbf{p}_m^r + \gamma \nabla \mathbf{L}_m^r))} \\ \mathbf{q}_{m,*}^{r+1} = \dfrac{\mathbf{q}_{m,*}^r + \gamma\theta_1(\mathbf{L}_m^r - \mathbf{L}_{m,*}^r)}{\max(1, \mathbf{abs}(\mathbf{q}_{m,*}^r + \gamma\theta_1(\mathbf{L}_m^r - \mathbf{L}_{m,*}^r))} \\ \mathbf{L}_m^{r+1} = \arg\min\limits_{\mathbf{L}_m} \sum\limits_i \theta_3 \sum\limits_{\partial_*} \|\partial_* \mathbf{A}_m \mathbf{L}_m - \partial_* \mathbf{B}_m\|_2^2 + \\ \dfrac{\left\|[\mathbf{L}_m - \eta((\nabla \mathbf{p}_m^{r+1})^T + \theta_1(\mathbf{q}_{m,+*}^{r+1} - \mathbf{q}_{m,-*}^{r+1})^T)] - \mathbf{L}_m^r\right\|^2}{2\eta} \end{cases} \tag{13}$$

where $\mathbf{p}_m$, $\mathbf{q}_{m,*}$ are the dual variables, γ and η are the step variants which can be modified at each iteration, and r is the iteration number.

5. Experiments

To demonstrate the effectiveness of our method, we evaluate it on two datasets: the synthetic chair sequence [31] and KITTI dataset [6]. We discuss our results on both datasets in the following sections.

5.1. Experimental Setup

Initialization. Our model is formulated on three consecutive stereo pairs. In particular, we treat the middle frame in the left view as the reference image. We adopt the StereoSLIC [41] to generate the superpixels. Given the stereo images, we apply the approach in [8] to obtain sparse feature correspondences. The traditional SGM [13] method is applied to obtain a disparity map which is used to initialize the plane parameters. The motion hypotheses are generated using RANSAC as implemented in [8]. In order to obtain the model parameters $\{\theta\}$ and $\{\alpha\}$, we performed block-coordinate-descent on a subset of 30 randomly selected training images.

Evaluations. Since our method estimates the scene flow and deblurs the images, we evaluate these two tasks separately. For the scene flow estimation results, we evaluate both the optical flow and disparity map by the same error metric, which is by counting the number of pixels having errors more than 3 pixels and 5% of its ground-truth. We adopt the PSNR to evaluate the deblurred image sequences for left and right view separately. Thus, for each sequence, we report three values: disparity errors for three stereo image pairs, flow errors in forward and backward directions, and PSNR values for six images.

Baseline Methods. As for our scene flow results, we compare with piece-wise rigid scene flow method (PRSF) [36], which ranks the first on KITTI stereo and optical flow benchmark. Note that PRSF is used as a preprocessing stage in [31]. We then compare our deblurring results with the state-of-the-art deblurring approach for monocular video sequence [16], and the approach for stereo videos [31].

Table 1. Quantitative comparisons on the Blurred KITTI dataset.

KITTI Dataset		Disparity		Flow		PSNR	
		m	m+1	Left	Right	Left	Right
Vogal *et al*. [36]		8.20	8.50	13.62	14.59	/	/
Kim and Lee [16]		/	/	38.89	39.45	28.25	29.00
Sellent *et al*. [31]		8.20	8.50	13.62	14.59	27.75	28.52
Ours	2 Frames	7.02	8.55	11.44	19.34	**30.24**	**30.71**
	3 Frames	**6.82**	**8.36**	**10.01**	**11.45**	29.80	30.30

5.2. Experimental Results

Results on KITTI. To the best of our knowledge, there are no realistic benchmark datasets that provide blur and its corresponding ground-truth clear images and scene flow. In this paper, we take advantage of the KITTI dataset [6] to create a synthetic **Blurred KITTI dataset** (will be publicly available) on realistic scenery. It contains 199 scenes, each of which includes 6 images of size 375×1242. Since the KITTI benchmark does not provide dense ground-truth flow, we use a state-of-the-art scene flow method [25] to generate dense ground-truth flows. Given the dense scene flow, the blur images are generated by using the piecewise linear 2D kernel, please refer to [16] and [31] for more details. The blur is caused by both objects motion and camera motion with occlusion and shadow.

We evaluated results by averaging errors and PSNR scores over $m-1$ to $m+1$ stereo image pairs. Table 1 shows the PSNR values, disparity errors, and flow errors averaged over the Blurred KITTI dataset. Our method consistently outperforms all baselines. We achieve the minimum error scores of 10.01% for optical flow and 6.82% for disparity in the reference view. In Fig. 5, we show qualitative results of our method and other methods on sample sequences from our dataset. Fig. 6 and Fig. 7 show the scene flow estimation and deblurring results of the Blurred KITTI dataset.

We then choose a subset of 50 more challenging sequences with large motion from the 199 scenes as test images, which contains daily traffic scenes covering urban areas (30 sequences), rural areas (10 sequences) and highway (10 sequences). Table 2 shows the PSNR values, disparity errors, and flow errors averaged over 50 test sequences on Blurred KITTI dataset. Fig. 8 (left) shows the performance of our deblurring stage with respect to the number of iterations. While we use 5 iterations for all our experiments, our experiments indicate that only 3 iterations are sufficient in most cases to achieve optimal performance under our model.

Results on Sellent *et al*. [31] dataset We further evaluate our approach on the dataset in [31] where the blur images are generated by 3D kernel model. Those sequences contain four real and four synthetic scenes and each of them includes six blur images with its sharp images, where ground-

(a) Blur Image (b) Kim and Lee [16] (c) Sellent *et al*. [31] (d) Ours

Figure 5. Numerous outdoor blurry frames and our deblurring result compare with several baselines. Best Viewed on Screen.

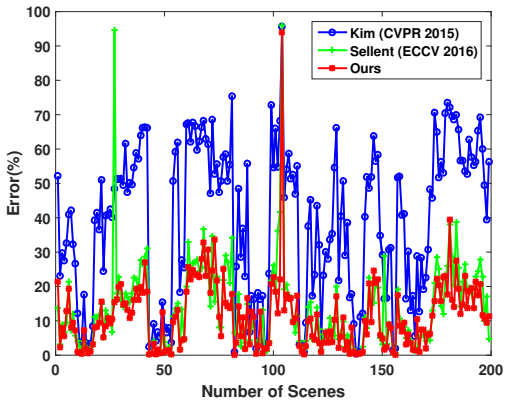


Figure 6. Flow estimation errors on the Blurred KITTI dataset. Our method clearly outperforms both monocular and stereo video deblurring methods.

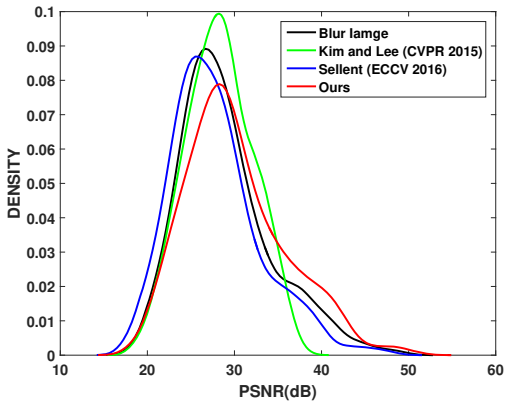


Figure 7. The distribution of the PSNR scores on the Blurred KITTI dataset. The probability distribution function for each PSNR was estimated using kernel density estimation with a normal kernel function. The heavy tail of our method means larger PSNR can be achieved using our method.

truth scene flow is only available for the synthetic scene "Chair". We thus report the quantitative comparison in Table 3 on the scene "Chair" between our method and state-of-the-art methods, where the evaluation results are averaged over 4 images. We also present the qualitative results in Fig. 9 for real images in this dataset. Fig. 8 (right)

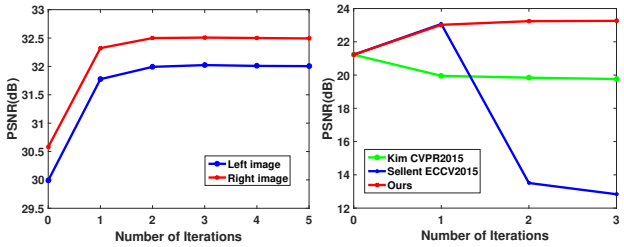
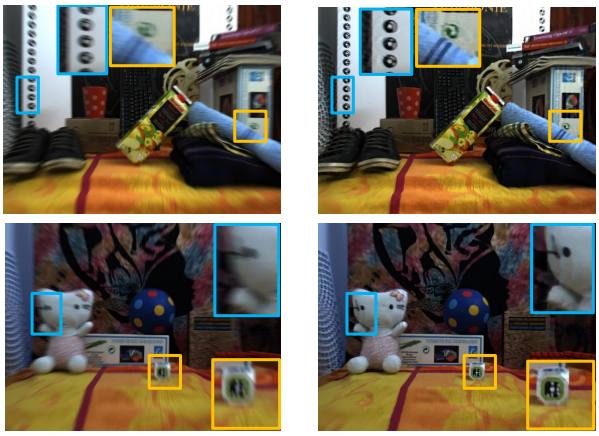


Figure 8. Deblurring performance with respect to iterations. (left) Our method gains an improvement of 0.3dB between the first and the last iteration on the 50 challenging dataset. (right) Comparison between our method and other baselines on the 'Chair' sequence.

Table 2. Quantitative comparisons on 50 challenging sequences.

Our Dataset		Disparity		Flow		PSNR	
		m	m+1	Left	Right	Left	Right
Vogal *et al*. [36]		6.67	6.70	7.26	7.90	/	/
Kim and Lee [16]		/	/	25.83	26.36	29.58	30.30
Sellent *et al*. [31]		6.67	6.70	7.26	7.90	28.73	29.44
Ours	2 Frames	4.98	5.82	**6.12**	13.06	**32.22**	**32.62**
	3 Frames	**4.90**	**5.76**	6.16	**6.17**	31.80	32.28

shows the performance comparison in deblurring between our method and other baselines with respect to iterations on scene "Chair". These results affirm our assumption that simultaneously solving scene flow and video deblur benefit each other and that a simple combination of two stages cannot achieve the targeted results.



(a) Blur Images (b) Our results

Figure 9. Sample deblur results on the real image dataset from Sellent et al.[31]. Best Viewed on Screen.

Results on another blur model: We have also tested our method on another blur generation model, where the blurred image is an average of consecutive three frames [9, 27]. The results are shown in Table 4 and Fig. 10 respectively, where our method again achieves the best performance.

Runtime: In all experiments, we simultaneously compute two direction scene flow and restoration six blur images.

Table 3. Performance comparisons on scene "Chair" [31].

Chair video		Disparity(%)	Flow Error(%)	PSNR(dB)
Menze [25]		1.17	9.33	/
Vogel [36]		1.34	2.13	/
Kim [16]		/	9.08	19.95
Sellent [31]		1.34	2.13	23.07
Ours	2 Frames	1.28	1.22	23.13
	3 Frames	**1.15**	**1.18**	**23.26**

Table 4. Quantitative evaluation on the KITTI dataset where the blur images are generated by averaging three consecutive frames.

	Kim [16]	Sellent *et al*. [31]	Ours
PSNR(dB)	23.21	23.31	**23.89**
SSIM	0.781	0.764	**0.786**

(a) Blur Image (b) Kim and Lee [16]

((c) Sellent *et al*. [31] (d) Ours

Figure 10. Quantitative evaluation on the KITTI dataset, where the blur images are generated by averaging three consecutive frames.

Our MATLAB implementation with C++ wrappers requires a total runtime of 40 minutes for processing one scene(6 images, 3 iterations) on a single i7 core running at 3.6 GHz.

6. Conclusion

In this paper, we present a joint optimization framework to tackle the challenging task of stereo video deblurring where scene flow estimation and video deblurring are solved in a coupled manner. Under our formulation, the motion cues from scene flow estimation and blur information could reinforce each other, and produce superior results than conventional scene flow estimation or stereo deblurring methods. We have demonstrated the benefits our framework on extensive synthetic and real stereo sequences.

Acknowledgement

This work was supported in part by China Scholarship Council (201506290130), Australian Research Council (ARC) grants (DP150104645, DE140100180), and Natural Science Foundation of China (61420106007, 61473230, 61135001), and Aviation fund of China (2014ZC5303). We thank all reviewers for their valuable comments.

References

[1] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011. 6

[2] Sunghyun Cho and Seungyong Lee. Fast motion deblurring. In *ACM Transactions on Graphics*, volume 28, pages 145:1–145:8. ACM, 2009. 2

[3] Sunghyun Cho, Jue Wang, and Seungyong Lee. Video deblurring for hand-held cameras using patch-based synthesis. *ACM Transactions on Graphics*, 31(4):64, 2012. 3

[4] Xiaoyu Deng, Yan Shen, Mingli Song, Dacheng Tao, Jiajun Bu, and Chun Chen. Video-based non-uniform object motion blur estimation and deblurring. *Neurocomputing*, 86:170–178, 2012. 1, 3

[5] Uwe Franke and Armin Joos. Real-time stereo vision for urban traffic scene understanding. In *IEEE Symposium on Intelligent Vehicles*, pages 273–278, 2000. 1

[6] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, page 0278364913491297, 2013. 6

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3354–3361, 2012. 1, 2

[8] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *IEEE Symposium on Intelligent Vehicles*, pages 963–968, 2011. 6

[9] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton van den Hengel, and Qinfeng Shi. From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur. *arXiv preprint arXiv:1612.02583*, 2016. 8

[10] Ankit Gupta, Neel Joshi, C Lawrence Zitnick, Michael Cohen, and Brian Curless. Single image deblurring using motion density functions. In *Proc. Eur. Conf. Comp. Vis.*, pages 171–184. Springer, 2010. 1, 2

[11] XC He, T Luo, SC Yuk, KP Chow, K-YK Wong, and RHY Chung. Motion estimation method for blurred videos and application of deblurring with spatially varying blur kernels. In *IEEE International Conference on Computer Sciences and Convergence Information Technology (ICCIT)*, pages 355–359, 2010. 3

[12] Michael Hirsch, Christian J Schuler, Stefan Harmeling, and Bernhard Schölkopf. Fast removal of non-uniform camera shake. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 463–470, 2011. 2

[13] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):328–341, 2008. 6

[14] Zhe Hu, Li Xu, and Ming-Hsuan Yang. Joint depth estimation and camera shake removal from single blurry image. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2893–2900, 2014. 2

[15] Tae Hyun Kim and Kyoung Mu Lee. Segmentation-free dynamic scene deblurring. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2766–2773, 2014. 3

[16] Tae Hyun Kim and Kyoung Mu Lee. Generalized video deblurring for dynamic scenes. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5426–5434, 2015. 1, 2, 3, 4, 6, 7, 8

[17] Jiaya Jia. Mathematical models and practical solvers for uniform motion deblurring. *Motion Deblurring: Algorithms and Systems*, page 1, 2014. 1, 2

[18] Sing Bing Kang. Automatic removal of chromatic aberration from a single image. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1–8, 2007. 1

[19] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-laplacian priors. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 1033–1041, 2009. 5

[20] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 233–240, 2011. 5

[21] Seungyong Lee and Sunghyun Cho. Recent advances in image deblurring. In *SIGGRAPH Asia Courses*, page 6. ACM, 2013. 1

[22] Feng Li, Jingyi Yu, and Jinxiang Chai. A hybrid camera for motion deblurring and depth map super-resolution. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1–8, 2008. 3

[23] Xin Li. Fine-granularity and spatially-adaptive regularization for projection-based image deblurring. *IEEE Trans. Image Proc.*, 20(4):971–983, 2011. 1

[24] Yasuyuki Matsushita, Eyal Ofek, Weina Ge, Xiaoou Tang, and Heung-Yeung Shum. Full-frame video stabilization with motion inpainting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(7):1150–1163, 2006. 3

[25] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3061–3070, 2015. 2, 3, 4, 5, 6, 8

[26] Tomer Michaeli and Michal Irani. Blind deblurring using internal patch recurrence. In *Proc. Eur. Conf. Comp. Vis.*, pages 783–798. Springer, 2014. 2

[27] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. *arXiv preprint arXiv:1612.02177*, 2016. 8

[28] SK Nayar and M Ben-Ezra. Motion-based motion deblurring. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(6):689–698, 2004. 3

[29] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comp. Vis.*, 47(1-3):7–42, 2002. 2

[30] Christian J Schuler, Michael Hirsch, Stefan Harmeling, and Bernhard Schölkopf. Blind correction of optical aberrations. In *Proc. Eur. Conf. Comp. Vis.*, pages 187–200. Springer, 2012. 1

[31] Anita Sellent, Carsten Rother, and Stefan Roth. Stereo video deblurring. In *Proc. Eur. Conf. Comp. Vis.*, pages 558–575. Springer, 2016. 1, 2, 3, 4, 6, 7, 8

[32] Hee Seok Lee and Kuoung Mu Lee. Dense 3d reconstruction from severely blurred images using a single moving camera. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 273–280, 2013. 3

[33] Jianping Shi, Li Xu, and Jiaya Jia. Just noticeable defocus blur detection and estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 657–665, 2015. 1

[34] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 769–777, 2015. 1, 2

[35] Yu-Wing Tai, Hao Du, Michael S Brown, and Stephen Lin. Image/video deblurring using a hybrid camera. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1–8, 2008. 3

[36] Christoph Vogel, Konrad Schindler, and Stefan Roth. 3d scene flow estimation with a piecewise rigid scene model. *Int. J. Comp. Vis.*, 115(1):1–28, 2015. 6, 8

[37] Oliver Whyte, Josef Sivic, Andrew Zisserman, and Jean Ponce. Non-uniform deblurring for shaken images. *Int. J. Comp. Vis.*, 98(2):168–186, 2012. 3

[38] Jonas Wulff and Michael Julian Black. Modeling blurred video with layers. In *Proc. Eur. Conf. Comp. Vis.*, pages 236–252. Springer, 2014. 3

[39] Li Xu and Jiaya Jia. Depth-aware motion deblurring. In *Proc. IEEE Int. Conf. Computational Photography*, pages 1–8, 2012. 3

[40] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1107–1114, 2013. 2

[41] Koichiro Yamaguchi, David McAllester, and Raquel Urtasun. Robust monocular epipolar flow estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1862–1869, 2013. 3, 6

[42] Shicheng Zheng, Li Xu, and Jiaya Jia. Forward motion deblurring. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1465–1472, December 2013. 1