This CVPR2015 extended abstract is the Open Access version, provided by the Computer Vision Foundation.

Simultaneous Video Defogging and Stereo Reconstruction

Zhuwen Li[1], Ping Tan[2], Robby T. Tan[3], Danping Zou[4], Steven Zhiying Zhou[1,5], Loong-Fah Cheong[1]
[1]National University of Singapore. [2]Simon Fraser University. [3]SIM University. [4]Shanghai Jiao Tong University. [5]NUS (Suzhou) Research Institute.

Fog generally poses challenges for multi-view stereo (MVS) algorithms, however it also importantly contains depth cues that are qualitatively different from stereo vision. Moreover, MVS actually helps resolve the airlight-albedo ambiguity in defogging. In this paper, we introduce a method to jointly estimate scene depth and recover the clear latent image from a foggy video sequence. In our formulation, the depth cues from stereo matching and fog information reinforce each other, producing superior results than conventional stereo or defogging algorithms.

A widely used fog scattering model is [2]:

$$I(\mathbf{x}) = J(\mathbf{x})\alpha(\mathbf{x}) + A(1-\alpha(\mathbf{x})), \tag{1}$$

where I is the observed image in scattering media, J is the latent clear image, A is the global atmospheric light, and α is the medium transmission determining the portion of the light that is not scattered and reaches the camera. When the atmosphere is homogeneous, the transmission α can be expressed as $\alpha(\mathbf{x}) = e^{-\beta z(\mathbf{x})}$, where β is the scattering coefficient depending on the density of the media, and z is the distance from the scene point to the camera center. To simplify the formulation, we assume that the scene point depth can approximate z well as in [1].

To formulate the problem of video-based stereo reconstruction, we assume n continuous frames $\mathcal{I} = \{I_t | t = 1,\ldots,n\}$ with known camera parameters $\mathcal{C} = \{\mathbf{K}_t, \mathbf{R}_t, \mathbf{t}_t | t = 1,\ldots,n\}$. We follow [6] to estimate the inverse depth maps $\mathcal{D} = \{D_t | t = 1,\ldots,n\}$ for all the frames. That is, $D_t(\mathbf{x}) = 1/Z_t(\mathbf{x})$, and $Z_t(\mathbf{x})$ is the depth of pixel $\mathbf{x}$ in frame t. To formulate the problem into a generic random field for dense image labeling, the continuous value of D_t is discretized into equal steps within some range $[d_{\min}, d_{\max}]$. The energy function then takes the following form:

$$E(\mathcal{D}) = \sum_{t=1}^{n}(E_p(D_t) + \eta E_g(D_t) + \rho E_s(D_t)), \tag{2}$$

where $E_p(D_t)$ is the photoconsistency term, $E_g(D_t)$ is the geometric coherence term and $E_s(D_t)$ is the smoothness term.

$E_p(D_t)$ measures the photoconsistency of frame t and its neighboring frames and is defined as

$$E_p(D_t) = \frac{1}{|\mathcal{N}(t)|}\sum_{t'\in\mathcal{N}(t)}\sum_{\mathbf{x}}\|I_t(\mathbf{x}) - I_{t'}(l_{t\to t'}(\mathbf{x}, D_t(\mathbf{x})))\|, \tag{3}$$

where $\mathcal{N}(t)$ denotes the neighboring frames of t and $l_{i\to j}(\mathbf{x}, D_i(\mathbf{x}))$ projects the pixel $\mathbf{x}$ with inverse depth $D_i(\mathbf{x})$ in frame i to frame j. However, this measurement becomes less accurate in a foggy video, because the scene radiance is attenuated differently from different camera positions. To overcome this difficulty, we take the scattering effect into consideration and define the new photoconsistency term that is corrected for scattering effect:

$$E_{ps}(D_t) = \frac{1}{|\mathcal{N}(t)|}\sum_{t'\in\mathcal{N}(t)}\sum_{\mathbf{x}}\|\hat{I}_{t'}(\mathbf{x}) - I_{t'}(l_{t\to t'}(\mathbf{x}, D_t(\mathbf{x})))\|, \tag{4}$$

where $\hat{I}_{t'}(\mathbf{x}) = (I_t(\mathbf{x}) - A)\frac{\pi_{t\to t'}(\mathbf{x},\alpha_t(\mathbf{x}))}{\alpha_t(\mathbf{x})} + A$ and $\pi_{i\to j}(\mathbf{x}, \alpha_i(\mathbf{x}))$ computes the corresponding transmission value in the j-th frame for the pixel $\mathbf{x}$ in the i-th frame with transmission $\alpha_i(\mathbf{x})$. Computing $\hat{I}_{t'}(\mathbf{x})$ can be interpreted as synthesizing the attenuated appearance of pixel of $\mathbf{x}$ in the t' frame with given transmission $\alpha_t(\mathbf{x})$. Note that $\alpha_t(\mathbf{x})$ can be related to $D_t(x)$ from $\alpha(\mathbf{x}) = e^{-\beta z(\mathbf{x})}$, so D_t is the only unknown in Equation (4).

Figure 1 (b)(c) show the values of our improved data term at the two points marked in Figure 1(a). Since these faraway points are highly attenuated and thus suffer from low image contrast, the conventional data term does not work and tends to assign incorrect depth values to these points. In comparison, the new photoconsistency cost shows a clear minimum at the position of the true inverse depth.

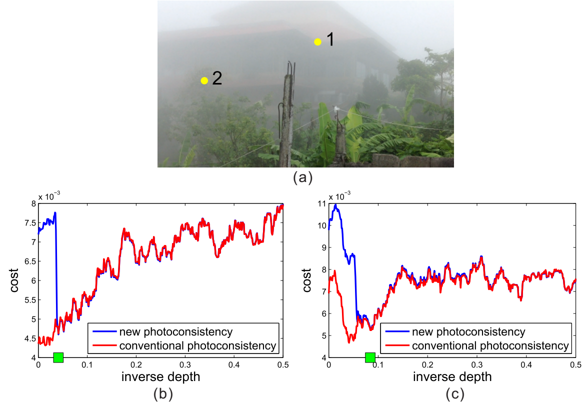

Figure 1: The new photoconsistency term: (a) A source frame from the "Bali" data with two heavily attenuated pixels, (b) The data cost at pixel 1. (c) The data cost at pixel 2. The green squares mark the true inverse depth (manually verified by projecting to other frames).

Meanwhile, the presence of transmission in Equation (4) opens up the possibility of enriching the details of the reconstructed depth, because a fog transmission map satisfies the Laplacian smoothness prior [4]. Concerning this, we find that this prior not only refines the transmission map, but also helps to preserve details in the depth map, probably due to its close relation to spectral image segmentation. Therefore, we add a Laplacian term

$$E_{Lap}(D_t) = vec(\alpha_t)^T \mathbf{L}_t vec(\alpha_t), \tag{5}$$

where $vec(\alpha_t)$ converts α_t into vector form, and $\mathbf{L}_t$ is the matting Laplacian matrix [5].

We also find that the fog transmission conveys more reliable constraint on depth order between points than on their absolute depth values. We further leverage on this aspect of fog information. More specifically, assume $\mathbf{x}$ and $\mathbf{y}$ are two neighboring pixels. If $\alpha_t(\mathbf{x}) > \alpha_t(\mathbf{y})$, we expect $D_t(\mathbf{x}) \geq D_t(\mathbf{y})$. Thus, when it is violated, we assign a large penalty τ_2. Since this condition also encodes the pairwise neighboring relationships, it is easy to incorporate it into $E_s(D_t)$, resulting in a smoothness term with ordering constraint $E_{so}(D_t)$ (specific formulation can be found in the paper).

Finally, our new energy function takes the following form

$$E(\mathcal{D}) = \sum_{t=1}^{n}(E_{ps}(D_t) + \eta E_g(D_t) + \rho E_{so}(D_t) + \lambda E_{Lap}(D_t)). \tag{6}$$

To solve this problem, we adopt an alternating optimization strategy with half quadratic splitting [3], based on the idea of introducing an auxiliary variable to decouple the terms and update them alternatingly. More details are described in the paper

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

[1] Laurent Caraffa and Jean-Philippe Tarel. Stereo reconstruction and contrast restoration in daytime fog. In *Proc. ACCV*. 2012.

[2] Fabio Gagliardi Cozman and Eric Krotkov. Depth from scattering. In *Proc. CVPR*, 1997.

[3] Donald Geman and George Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14 (3):367–383, 1992.

[4] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33 (12):2341–2353, 2011.

[5] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30 (2):228–242, 2008.

[6] Guofeng Zhang, Jiaya Jia, Tien-Tsin Wong, and Hujun Bao. Consistent depth maps recovery from a video sequence. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(6):974–988, 2009.