Simultaneous Video Defogging and Stereo Reconstruction

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

Abstract

We present 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, and produce superior results than conventional stereo or defogging algorithms. We first improve the photo-consistency term to explicitly model the appearance change due to the scattering effects. The prior matting Laplacian constraint on fog transmission imposes a detail-preserving smoothness constraint on the scene depth. We further enforce the ordering consistency between scene depth and fog transmission at neighboring points. These novel constraints are formulated together in an MRF framework, which is optimized iteratively by introducing auxiliary variables. The experiment results on real videos demonstrate the strength of our method.

1. Introduction

Multi-view stereo receives intensive investigations [32]. Most of these methods, however, are designed for images captured in clear scenes, and consequently foggy or underwater images present a significant challenge. One of the reasons is because these images are associated with poor image contrast, which causes image features to be less distinctive and confuses stereo matching. Another reason is that the photo-consistency measure in conventional stereo algorithms does not consider the scattering and absorption phenomenon during light propagation, and thus generate systematic matching errors.

While fog poses challenges for stereo algorithm, it also brings compensatory advantages. From a computational viewpoint, it is well known [16] that stereo does not work well for large distances. The depth smoothness prior further results in loss of surface details [42] such as thin elongated structures and holes. Fog transmission information (i.e. the α-channel) contains depth cues that are qualitatively different, because: 1) fog transmission provides depth ordering, since thicker fog is associated with larger distance, 2) fog transmission satisfies the matting Laplacian [21], which provides a detail preserving smoothness prior on scene depth.

At the same time, stereo vision helps defogging. Previous defogging techniques are mainly designed to defog a single image, which is essentially an ill posed problem due to the airlight-albedo ambiguity [10]. Recent techniques rely on additional heuristic assumptions, such as piecewise constant albedo [10], maximum local contrast [35] or dark channel prior [17]. But all these assumptions can be fooled by the airlight-albedo ambiguity. As a result, nearby objects with unsaturated color are often mistaken as faraway objects with saturated color. It is demonstrated [19] that even rough depth information can significantly improve the defogging performance. A precise depth estimation from stereo will help to reduce the airlight-albedo ambiguity in defogging.

A naive solution for simultaneous stereo and defogging is to apply both algorithms iteratively, e.g. iteratively defog all images and then apply stereo vision to the defogged results. However, most of existing defogging algorithms are designed for a single image, and will generate temporally inconsistent results when they are process on a frame-by-frame manner. Thus, the following stereo algorithms will generate large error on these inconsistently defogged images. This problem is shown in our experiments.

In this paper, we study stereo vision and defogging problems jointly, and design an algorithm that simultaneously estimates scene depth and defogs the input images. Our method is based on the observation that the depth cues from stereo and fog thickness are complementary to each other (i.e. stereo cues are more reliable for nearby objects and fog thickness cues are more reliable for faraway objects). Our method performs best on scenes with thick fog or large camera movement when both depth cues are strong at close and far. This feature makes our algorithm especially suitable for applications in autonomous navigation of unmanned vehicles in bad weather or underwater, such as in [23, 29].

Our method includes four key features. Firstly, we improve the photo-consistency term in stereo matching to incorporate the scattering effect. When evaluating the consistency of two pixels from different viewpoints, we explicitly model their appearance change due to fog. Secondly, we

*D. Zou worked on this project as a research fellow with P. Tan.

compute the fog transmission at each pixel directly from the scene depth (and the estimated fog density). This ensures the results from our stereo and defogging are consistent with each other, relieves the airlight-albedo ambiguity in defogging, and maintains temporal consistency in the final defogged video. Thirdly, we incorporate a strong prior on fog transmission into the joint stereo-defogging formulation. Specifically, we impose the matting Laplacian constraint [21] to the scene depth, since the fog transmission can be directly computed from depth. As we will see from the experiments, this constraint helps to capture the fine details in the depth map. Lastly, we also incorporate pairwise depth ordering constraint, leveraging on the reliability of fog transmission in conveying ordinal depth information.

2. Related Work

Many multi-view stereo (MVS) vision algorithms have been proposed and we summarize a few of the most recent and related works here. More detailed discussion of various stereo algorithms refers to the excellent survey [32], and the multi-view stereo evaluation for the recent top performing methods refers to [1]. In particular, some of the recent works that are more related to ours are those based on the multiple depth map approach [5, 6, 12, 13, 15, 34, 43].

A significant advantage that MVS algorithms have over two-view stereo is their ability to reason in a principled manner about visibility and occlusions. MVS algorithms can also impose the consistency between disparity estimates at different frames to reduce sensitivity to outliers in individual frames. Similar to the work of [43], we add a geometric coherence term to the conventional data term based on photoconsistency, to impose all the preceding geometric constraints. Despite the improved results, all these energy functions, when minimized by global optimization techniques such as multi-label graph-cut method [4], still suffer from characteristic artifacts. The resultant depth maps typically exhibit shrinking bias (shortcutting of the segmentation boundary due to bias toward shorter boundaries). Irregular or thin objects such as trees, bushes, branches, fences are often poorly reconstructed or even completely missed (see the examples in [13], or the flower sequence in [43]). The reconstruction results also depend on the evolving quality of the current depth estimate, which in the case of a foggy or underwater scene might be too poor to yield good results. Our work not only incorporates the scattering effect to improve stereo matching, it also makes use of the fog transmission information to preserve details in depth maps.

A number of single-image dehazing or defogging methods have been proposed [2, 10, 17, 24, 28, 35, 36]. Tan's method [35] obtains airlight, i.e. light scattered by particles, by maximizing local contrast. Fattal's [10] decomposes shading and transmission functions by assuming both of them are statistically uncorrelated locally. He et al.'s [17] introduces a dark channel prior that exploits the minimum intensity of all color channels in a local window to indicate the level of haziness. The most recent work, Meng et al.'s [24], introduces a boundary constraint on the transmission function. Ancuti et al.'s [2], instead of dealing with fog, turns to underwater vision by employing image fusion. While all these single image defogging methods work to some extent, their major problem is the ambiguity between airlight and albedo. Nearby objects with low color saturation will be considered to be far away objects with saturated color, since the methods will mistakenly think the surfaces are under heavy airlight. Aside from using a single image, a few methods have been introduced to utilize multiple images [25, 31, 38]. These methods require pixelwise registration of images with different particle data. Consequently, they cannot be applied if the scenes are dynamic, or if the camera moves.

Recently, several methods are introduced [7, 27, 30] to solve stereo for foggy or underwater images. All of them are designed for a pair of stereo images, while our method processes a video sequence from a monocular moving camera. Caraffa and Tarel's method [7] combines the conventional photo-consistency term and the scattering equation to simultaneously solve stereo and defog. However, as we discussed in Section 4.1, the conventional photo-consistency term becomes less effective in scattering media. Furthermore, the scattering equation is sensitive to the nonlinear camera response function and image noise. As a result, this method is only demonstrated on synthetic data. Nascimentoet al. [27] and Roser et al.'s method [30] iterates two steps: applying a conventional stereo algorithm to compute the dense depth; taking the depth to inverse the scattering equation to estimate the clear latent image. They further apply the matting Laplacian [21] as an edge preserving filter to enhance the depth in the iteration. This straightforward combination of suffers from two flaws. Firstly, the light scattering effects is not modeled in the photo-consistency measure for stereo matching, which causes erroneous stereo reconstruction. Secondly, different video frames are defogged independently, which generates inconsistent frames and leads to systematic error in stereo reconstruction. Our experiments verify these problems. In comparison, our formulation presents a tight fusion of the depth cues from stereo and defogging and produces much stronger results.

Our work is also related to [26] which studies structured light based stereo in scattering media. In contrast to this, however, our work focuses on passive stereo in scattering media with completely different formulation and setup.

3. Background

Before formulating our approach, we give an overview of the typical formulations for the defogging and MVS problems and introduce the notations used in the paper.

3.1. Fog model

In computer vision, a widely used scattering model is [9, 10, 17, 25, 35]:

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

where I is the observed image in scattering media (e.g. fog, haze, or turbid water), J is the latent clear image unaffected by the media, 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:

$$\alpha(\mathbf{x}) = e^{-\beta z(\mathbf{x})}, \quad (2)$$

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 [7, 19].

3.2. Stereo from monocular videos

Assume n continuous frames $\mathcal{I} = \{I_t | t = 1, \ldots, n\}$ with known camera parameters $\mathcal{C} = \{\mathbf{K}_t, \mathbf{R}_t, \mathbf{t}_t | t = 1, \ldots, n\}$, where $\mathbf{K}_t$ is the intrinsic matrix, $\mathbf{R}_t$ is the rotation matrix and $\mathbf{t}_t$ is the translation vector. These camera parameters can be estimated by any standard structure-from-motion (SfM) methods [18, 40]. We follow [43] in formulating the problem of video-based stereo reconstruction, which aims 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:

$$E(\mathcal{D}) = \sum_{t=1}^{n} (E_p(D_t) + \eta E_g(D_t) + \rho E_s(D_t)), \quad (3)$$

where $E_p(D_t)$ is the photoconsistency term, $E_g(D_t)$ is the geometric coherence term, $E_s(D_t)$ is the smoothness term, and η and ρ are the parameters to balance these terms.

In order to define the photoconsistency term, we assume $\mathbf{x}$ is written in the homogeneous coordinate and derive from multi-view geometry the following projection function:

$$l_{i\to j}(\mathbf{x}, D_i(\mathbf{x})) = \mathbf{K}_j \mathbf{R}_j \mathbf{R}_i^T \mathbf{K}_i^{-1} \mathbf{x} + D_i(\mathbf{x}) \mathbf{K}_j (\mathbf{t}_j - \mathbf{R}_j \mathbf{R}_i^T \mathbf{t}_i), \quad (4)$$

which projects the pixel $\mathbf{x}$ with inverse depth $D_i(\mathbf{x})$ in frame i to frame j. Now we can write the photoconsistency term as

$$E_p(D_t) = \frac{1}{|\mathcal{N}(t)|} \sum_{t' \in \mathcal{N}(t)} \sum_{\mathbf{x}} \| I_t(\mathbf{x}) - I_{t'}(l_{t\to t'}(\mathbf{x}, D_t(\mathbf{x}))) \|, \quad (5)$$

where $\mathcal{N}(t)$ denotes the neighboring frames of t and $|\mathcal{N}(t)|$ is the number of frames in the neighboring set. As in many classic stereo algorithms, this term measures the photoconsistency of frame t and its neighboring frames.

The geometric coherence term is specifically designed for video-based stereo [43] to ensure temporal consistency of recovered depth maps and to handle occlusions. It is still a unary cost but it checks the inverse depths of the conjugate pixels in neighboring frames:

$$E_g(D_t) = \frac{1}{|\mathcal{N}(t)|} \sum_{t' \in \mathcal{N}(t)} \sum_{\mathbf{x}} \| \mathbf{x} - l_{t'\to t}(\mathbf{x}', D_{t'}(\mathbf{x}')) \|, \quad (6)$$

where $\mathbf{x}' = l_{t\to t'}(\mathbf{x}, D_t(\mathbf{x}))$ is the conjugate pixel location of $\mathbf{x}$ in frame t' and $D_{t'}$ is the inverse depth map of frame t'. This term essentially enforces the geometric consistency of depth maps of different frames, contributing to the temporal consistency of the final result.

Typically, the smoothness term is defined as:

$$E_s(D_t) = \sum_{\mathbf{x}} \sum_{\mathbf{y} \in \mathcal{N}(\mathbf{x})} w(\mathbf{x}, \mathbf{y}) \cdot f(D_t(\mathbf{x}), D_t(\mathbf{y})). \quad (7)$$

For robustness, $f(D_t(\mathbf{x}), D_t(\mathbf{y}))$ is usually the truncated ℓ_1 function:

$$f_{\ell_1}(D_t(\mathbf{x}), D_t(\mathbf{y})) = \min\{\|D_t(\mathbf{x}) - D_t(\mathbf{y})\|, \tau_1\} \quad (8)$$

where τ_1 is the truncating parameter and $w(\mathbf{x}, \mathbf{y})$ is the weight function indicating the probability that $\mathbf{x}$ and $\mathbf{y}$ should be assigned the same inverse depth. To encourage the depth discontinuity to be coincident with color change, $w(\mathbf{x}, \mathbf{y})$ is usually defined based on the color difference of neighboring pixels [3, 4, 33].

Since it is difficult to achieve global optimality for Equation (3), video-based stereo usually adopts a two-step strategy. Firstly, the depth initialization step solves Equation (3) frame by frame by omitting the geometric coherence term to obtain the initial depth maps. Then, the estimated initial depth maps are fed into the second step to solve the complete version of Equation (3). More specifically, it now solves one depth map at a time by fixing the depth maps of the other frames, iterating several passes (typically 2 passes) with one pass traversing all frames once. These problems are standard MRF minimization problems that admit efficient solutions like graph cut [4] or loopy belief propagation (LBP) [11]. For more details, interested readers can refer to [43].

4. Simultaneous Defogging and Stereo

It is often difficult to measure the photoconsistency in a foggy video, because the scene radiance is attenuated differently from different camera positions. To overcome this difficulty, we propose a more sophisticated photoconsistency term which takes the scattering effect into consideration. Meanwhile, the presence of fog also opens up the possibility of enriching the details of the reconstructed depth. For this purpose, we include the matting Laplacian [21] constraint as a detail preserving smoothness term. And lastly, we also

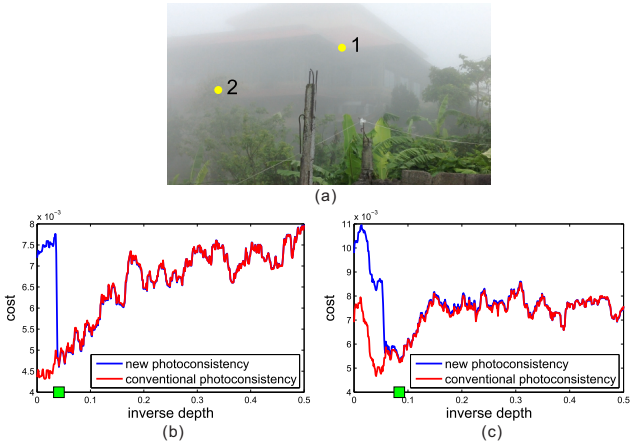

Figure 1. The new photoconsistency term: (a) A source frame from the "Bali" data with two heavily attenuated pixels, (b) The data cost at pixel 1. (c) The data cost at pixel 2. The green squares mark the true inverse depth (manually verified by projecting to other frames).

leverage on the ordering constraint provided by the transmission when imposing smoothness on the inverse depths. Consequently, our new energy function takes the following form

$$E(\mathcal{D}) = \sum_{t=1}^{n}(E_{ps}(D_t) + \eta E_g(D_t) + \rho E_{so}(D_t) + \lambda E_{Lap}(D_t)), \tag{9}$$

where $E_g(D_t)$ is the unchanged geometric coherence term, and $E_{ps}(D_t)$, $E_{so}(D_t)$ and $E_{Lap}(D_t)$ correspond to the new photoconsistency term, the new smoothness term augmented with an ordering constraint, and the matting Laplacian term, respectively. Jointly optimizing these terms helps bring mutual benefits to both the stereo depth recovery and defogging. In particular, ambiguity of faraway objects are lifted, and both depth and color details are better recovered. In the following subsections, we will explain the new energy terms in detail.

4.1. Photoconsistency

Currently let us assume that the scattering coefficient β and atmospheric light A are known. We will present an automatic method to estimate them in Section 4.5. Denote $\mathbf{R}_i = [\mathbf{r}_{i,1} \;\; \mathbf{r}_{i,2} \;\; \mathbf{r}_{i,3}]^T$ and $\mathbf{t}_i = [t_{i,1} \;\; t_{i,2} \;\; t_{i,3}]^T$, where $\{\mathbf{r}_{i,k}^T | k = 1, 2, 3\}$ are the rows of $\mathbf{R}_i$ and $\{t_{i,k} | k = 1, 2, 3\}$ are the entries of $\mathbf{t}_i$. Now we can derive the projection function $\pi_{i\to j}(\mathbf{x}, \alpha_i(\mathbf{x}))$ which computes the corresponding transmission value in the j-th frame for the pixel $\mathbf{x}$ in the i-th frame with transmission $\alpha_i(\mathbf{x})$. Specifically,

$$\pi_{i\to j}(\mathbf{x}, \alpha_i(\mathbf{x})) = \exp(\hat{\mathbf{r}}_{i,j}^T \mathbf{K}_i^{-1}\mathbf{x} \log(\alpha_i(\mathbf{x})) + \beta(\hat{\mathbf{r}}_{i,j}^T \mathbf{t}_i - t_{j,3})), \tag{10}$$

where $\hat{\mathbf{r}}_{i,j}^T = [\mathbf{r}_{j,3}^T\mathbf{r}_{i,1} \;\; \mathbf{r}_{j,3}^T\mathbf{r}_{i,2} \;\; \mathbf{r}_{j,3}^T\mathbf{r}_{i,3}]$ is the last row of $\mathbf{R}_j\mathbf{R}_i^T$. More intuitively, Equation (10) can be interpreted as follows. Knowing $\alpha_i(\mathbf{x})$ means knowing $D_i(\mathbf{x})$ because they can be converted to each other by Equation (2). Thus, together with the camera parameters, the 3D world coordinate $\mathbf{X}_{world}$ of pixel $\mathbf{x}$ can be calculated. With the world coordinate $\mathbf{X}_{world}$, we can calculate its depth in the j-th camera, and hence the transmission $\pi_{i\to j}(\mathbf{x}, \alpha_i(\mathbf{x}))$. Note that we use depth $Z_i(\mathbf{x})$ instead of distance $z_i(\mathbf{x})$ in Equation (2) to simplify Equation (10), though in principal the distance $z_i(\mathbf{x})$ can be used too.

Now we can define the new photoconsistency term that is corrected for scattering effect:

$$E_{ps}(D_t) = \frac{1}{|\mathcal{N}(t)|}\sum_{t'\in\mathcal{N}(t)}\sum_{\mathbf{x}} \|\hat{I}_{t'}(\mathbf{x}) - I_{t'}(l_{t\to t'}(\mathbf{x}, D_t(\mathbf{x})))\|, \tag{11}$$

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 computing it 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 Equation (2), so D_t is the only unknown in Equation (11). For the same reason, the following functions defined on α_t will also be considered as functions of D_t unless specifically pointed out.

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 (marked by a green square in Figure 1 (b) and (c)).

4.2. Laplacian smoothing

A fog transmission map should satisfy the Laplacian smoothness prior [17, 21]. 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.

The Laplacian term is defined as

$$E_{Lap}(D_t) = vec(\alpha_t)^T \mathbf{L}_t vec(\alpha_t), \tag{12}$$

where $vec(\alpha_t)$ converts α_t into vector form, and $\mathbf{L}_t$ is the Laplacian matrix with its (i, j)-th entry defined as [21]

$$\mathbf{L}_t(i,j) = \sum_{k|(i,j)\in w_k}(\delta_{ij} - \frac{1}{|w_k|}(1 + (I_t(\mathbf{x}_i) - \mu_k)(\Sigma_k + \frac{\varepsilon}{|w_k|}\mathbf{I}_3)^{-1}(I_t(\mathbf{x}_j) - \mu_k))) \tag{13}$$

where δ_{ij} is the Kronecker delta, μ_k is a 3×1 mean vector of the colors in a window w_k, Σ_k is a 3×3 covariance matrix of the colors in w_k, ε is a regularizing parameter, and $\mathbf{I}_3$ is the 3×3 identity matrix. Equation (13) sums over all the 3×3 windows w_k in which the i-th and j-th pixels both appear.

We demonstrate the effectiveness of this Laplacian s-

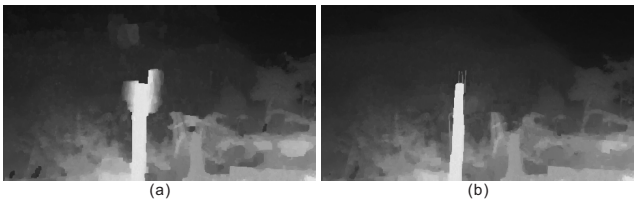

Figure 2. The Laplacian term: (a) and (b) are the inverse depth (of Figure 1 (a)) produced without and with the Laplacian term.

moothing in Figure 2. When the Laplacian term is not enabled, shape details on the nearby pole are missing, mostly because of the shrinking biased caused by the belief-propagation based optimization. In comparison, our Laplacian term can capture these details.

4.3. Ordering constraint based on transmission

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 $D_t(\mathbf{x}) \geq D_t(\mathbf{y})$. Theoretically, it is always true in Equation (9). However, it may disobey the ordering constraint, since we adopt an alternating optimization method (Section 4.4), which decouples $\alpha_t(\mathbf{x})$ and $D_t(\mathbf{x})$. Thus, when this condition is violated, we assign a large penalty τ_2. Mathematically, we modify $f(D_t(\mathbf{x}), D_t(\mathbf{y}))$ in Equation (7) as:

$$f_o(D_t(\mathbf{x}), D_t(\mathbf{y})) = \begin{cases} \tau_2 & \delta(\alpha_t(\mathbf{x}) - \alpha_t(\mathbf{y})) \cdot \delta(D_t(\mathbf{x}) - D_t(\mathbf{y})) = -1, \\ f_{\ell_1}(D_t(\mathbf{x}), D_t(\mathbf{y})) & \text{otherwise}, \end{cases} \tag{14}$$

where $\delta(\cdot)$ is the sign function that returns 1 for positive values, -1 for negative values and 0 for 0 values. Replacing f_{ℓ_1} with f_o in Equation (7), we denote the resultant smoothness term with ordering constraint as $E_{so}(D_t)$. Note that the weight function $w(\mathbf{x}, \mathbf{y})$ remains unchanged. It has been shown that when the transmission α_t is known, this smoothness function remains a metric [8] and is thus solvable by α-expansion.

Figure 3 shows the advantage of enforcing the ordering constraint. In Figure 3 (b), the inverse depth in the sky in the red box is wrong, while it is corrected in Figure 3 (c) by our ordering constraint. These results are produced by solving Equation (9), where the only difference is whether the smoothness term $E_{so}(D_t)$ is used.

4.4. Solver

Following [43], we also adopt a two-step optimization strategy. We initialize the depth maps by ignoring the geometric coherence term in the first step, and then solve the complete version of Equation (9) iteratively in the second step. Equation (9) is not easy to solve because of the Laplacian term. Thus, we adopt an alternating optimization strategy with half quadratic splitting [14], based on the idea of introducing an auxiliary variable to decouple the terms and update them alternatingly. This strategy is widely used in many computer vision algorithms, particularly in the total variation (TV) regularizations [39]. Although it originally solves convex problems and there is no general convergence theory applied to non-convex problems, it has been applied to several non-convex problems [20, 41] and shown to perform well. Indeed, empirically we found our algorithm also had strong convergence behaviour.

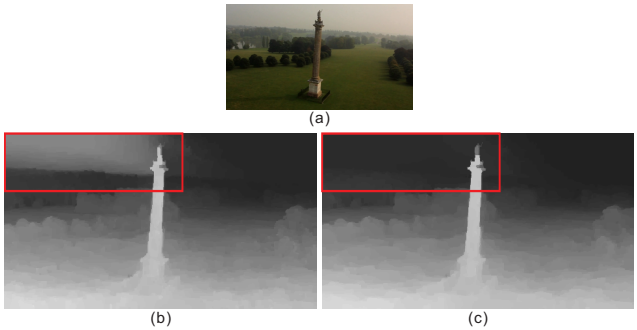

Figure 3. The ordering constraint: (a) A source frame from the "Blenheim" data. (b) and (c) are the inverse depths without and with the ordering constraint.

We split the function into two parts: one is a function of inverse depth, which can be minimized by MRF energy minimization; the other one is a function of transmission, which is a convex function and has a closed-form solution. We further introduce a coupling term to enforce the consistency of D_t and α_t. Thus the energy function is rewritten as

$$E(\mathcal{D}) = \sum_{t=1}^{n} (E_{ps}(D_t) + \eta E_g(D_t) + \rho E_{so}(D_t) + \epsilon \| e^{-\frac{\beta}{D_t}} - \alpha_t \|_F^2 + \lambda E_{Lap}(\alpha_t)). \tag{15}$$

We minimize the new objective function iteratively until convergence or the number of iteration exceeds the maximum limit. In each iteration, we solve for D_t while fixing α_t, and then solve for α_t while fixing D_t. Thus, the two subproblems are

$$\min \sum_{t=1}^{n} (E_{ps}(D_t) + \eta E_g(D_t) + \rho E_{so}(D_t) + \epsilon \| e^{-\frac{\beta}{D_t}} - \alpha_t \|_F^2), \tag{16}$$

$$\min \sum_{t=1}^{n} (\epsilon \| e^{-\frac{\beta}{D_t}} - \alpha_t \|_F^2 + \lambda E_{Lap}(\alpha_t)). \tag{17}$$

Since the last term in Equation (16) is unary, and $E_{so}(D_t)$ with the ordering constraint introduces no additional difficulty, Equation (16) can be solved by graph cut [4] or LBP [11]. The second subproblem Equation (17) is an unconstrained convex problem and has closed-form solution:

$$vec(\alpha_t^*) = (\mathbf{I} + \frac{\lambda}{\epsilon} \mathbf{L}_t)^{-1} \mathbf{u}_t, \tag{18}$$

where $\mathbf{u}_t = vec(e^{-\frac{\beta}{D_t}})$ and $\mathbf{I}$ is the identity matrix.

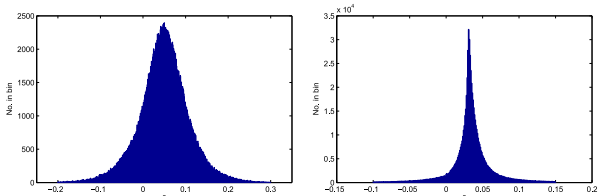

Figure 4. The histograms of β for: (a) Bali and (b) Synthetic.

4.5. Estimating A and β

Tan [35] uses the brightest pixel value as the atmospheric light A. Later the estimation of A is further refined in [10, 17]. We follow the method in [17] to estimate A because of its robustness.

We then estimate β, given A and a set of sparsely matched pixel pairs $\{(\mathbf{x}_k, \mathbf{y}_k)|k = 1, \ldots, K\}$ in the i-th and j-th frames. These points are readily collected from the structure-from-motion step where feature correspondences are established to recover camera motion. They satisfy the following equations according to Equation (1),

$$\begin{aligned} I_i(\mathbf{x}_k) - A &= (J_i(\mathbf{x}_k) - A)\alpha_i(\mathbf{x}_k), \\ I_j(\mathbf{y}_k) - A &= (J_j(\mathbf{y}_k) - A)\alpha_j(\mathbf{y}_k). \end{aligned} \tag{19}$$

Since $\mathbf{x}_k$ and $\mathbf{y}_k$ correspond to the same scene point, we know $J_i(\mathbf{x}_k) - A = J_j(\mathbf{y}_k) - A$ by photo-consistency. Then we have

$$\begin{aligned} \frac{I_i(\mathbf{x}_k) - A}{I_j(\mathbf{y}_k) - A} &= \frac{\alpha_i(\mathbf{x}_k)}{\alpha_j(\mathbf{y}_k)} = \frac{\exp(-\beta Z_i(\mathbf{x}_k))}{\exp(-\beta Z_j(\mathbf{y}_k))} \\ &= \exp(-\beta(Z_i(\mathbf{x}_k) - Z_j(\mathbf{y}_k))), \end{aligned} \tag{20}$$

where $Z_i(\mathbf{x}_k)$ is the depth (or distance) and readily known from the structure-from-motion step. Note that each pixel pair gives an estimation of β from Equation (20). For stability, we disregard those pairs where the inverse depth difference is smaller than some threshold (typically 10^{-3}). Thus, we build a histogram from the collection of β and choose the value of highest bin as β. Several examples of the histogram are shown in Figure 4, from which we can see the nice distribution of the estimations.

5. Experimental Results

In our experiments, the camera parameters for all frames in all data are estimated beforehand. Because of the scattering effect, we use SIFT detector and descriptor [22] to match feature points in videos. We find that most features are extracted from near objects, which are less degraded by scattering media, thus most matched features are confident. Then we use the SfM method proposed in [18] to recover camera poses. With the estimated camera poses, we first conduct a depth initialization step by solving Equation (9) without the geometric coherence term, and then solve the complete version of Equation (9) in the next step. The Equation (9) (with or without the geometric coherence term) is solved by iteratively solving Equation (16) and Equation (17). During the depth initialization step, we do not have the fog transmission map at our disposal yet, so the result of the dark channel method [17] is used as an initialization for α to apply the ordering constraint. In the subsequent iterations, the fog transmission map is kept updated from the estimated depth map via Equation (17).

We evaluate our method on several challenging videos. The videos are captured from different localities on foggy days. Among these videos, the "Blenheim" data is downloaded from Vimeo. We also captured a turbid underwater video in a tank using an underwater camera. To simulate the scattering medium, we poured milk in the water. Complete video results can be found in the supplementary material. All the experiments are run on a desktop with Intel quad-core 2.4GHz CPU. The time taken to process a frame with 480×270 image resolution is about 10 minutes.

For the stereo evaluation, we compare with the state-of-the-art video-based stereo reconstruction [43]. For the defogging evaluation, we compare to several state-of-the-art single image defogging methods, i.e. the dark channel method by He et al. [17], the latest algorithm by Meng et al. [24] and the improved version of [36], denoted as NBPC+PA [37]. For a more fair comparison, we also evaluate the approaches adopting both fog and stereo cues; these include our own implementation of Caraffa and Tarel (referred as CT) [7] and two straightforward combinations of defogging and stereo (similar to [27, 30]). The straightforward combinations simply iterate between defogging (by Meng et al. [24]) and stereo (by Zhang et al.[43]) with different initializations. We denote the one starting with the stereo result as SD (Stereo then defog) and the other one as DS (Defog then stereo). The iteration stops when there are no big changes in the results (usually 5 iterations).

We first verify our algorithm on a synthetic data. The evaluation metric is the error computed using the sum of the absolute difference between the recovered depth (or defogging image) and the groundtruth. The quantitative comparisons are shown in Table 1, from which it is observed that our approach outperforms the others with the smallest errors. In the following, we show the qualitative comparisons and analyze the strong aspects of our method.

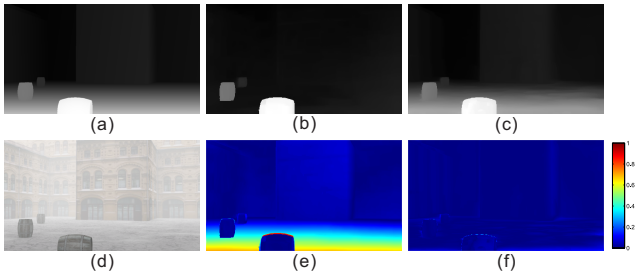

Figure 5. Comparison with conventional stereo method. (a) Groundtruth inverse depth. (b) Result from Zhang et al. [43]. (c) Our result. (d) Source image. (e) and (f) are the error maps of Zhang et al. [43] and ours respectively

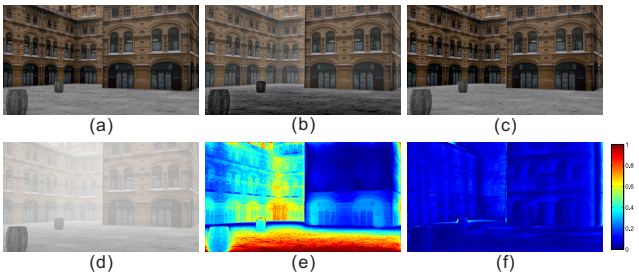

Figure 6. Comparison with conventional defogging methods. (a) Clear image. (b) Result from Meng et al. [24]. (c) Our result. (d) Source image. (e) and (f) are the error maps of Meng et al. [24] and ours respectively.

Figure 5 shows the evaluation of stereo estimation on the synthetic data. The estimated inverse depth map are normalized to $[0, 1]$ for comparison. As can be observed, our result is considerably close to the ground truth for both near and far regions, while Zhang et al. [43] cannot distinguish the depth of the ground and walls very well due to the low contrast. Thus, these regions are wrongly assigned close inverse depths.

For the defogging evaluation on the synthetic data, it is observed in Figure 6 that our method produces the closest result to the groundtruth according to the error map. In contrast, Meng et al. [24] (the results of other two single image defogging methods are similar) shows significant errors in the ground region, which is light-toned in color and is thus mistaken (due to the respective priors employed) as faraway objects densely covered with strong fog. The fog of the distant region in the middle of the image is not completely removed by these single image based method. Furthermore, single image defogging methods generate temporal flickering on videos, since each frame is processed independently. Please refer to the supplementary video.

We then compare our algorithm with the others on real-world data as shown in Figure 7. Comparing only to the stereo algorithm [43] (second column, every even row), it is observed from Figure 7 that [43] is not able to recover the depths of the distant objects, which are highly degraded by the scattering medium, such as the house in "Bali", the distant trees in "Blenheim" and the castle in "Underwater". In contrast, our method recovers the structures of the house, trees and castle more correctly. Moreover, [43] generates some artifacts in these data, such as the plant region in "Bali", the sky region in "Blenheim", the door region in "Motorcycle" and the region near the chest in "Underwater". In comparison, our method is free of such artifacts. Our method preserves the details of the scene, such as the elongated steel frame embedded in the foreground concrete column in "Bali", the more complicated geometrical details of trees and bushes in "Blenheim", and the furrows on the treasure chest in "Underwater".

Comparing only to the defogging algorithm [24] (second column, every odd row), it is observed from Figure 7 that our method is able to recover more faithful colors, particularly in the light-toned ground regions of the "motorcycle" and "playground" sequences, which again cause problems for single image defogging. Moreover, the fog in front of the right door of "Motorcycle" is not completely removed by [24] and recovered scene of "underwater" is darker. In contrast, our method handles these regions well thanks to the high quality depth estimated by our stereo algorithm.

Table 1. Comparison on the synthetic data. The error (per frame) of the stereo and defogging results are presented.

	[17]	[24]	[37]	[43]	CT[7]	DS	SD	our
Stereo	N/A	N/A	N/A	76.42	25.44	23.10	50.63	**8.59**
Defogging	29.42	31.68	33.98	N/A	11.92	8.32	14.36	**6.12**

Comparing to the two straightforward combinations of defogging and stereo, it is observed from Figure 7 that separate defogging and stereo does not necessarily improve each other much. For DS, though the initial defogged images look visually plausible (shown in the second column of Figure 7), they are actually inconsistent over frames and lead to poor stereo results. For instances, the stereo results of the region near the pillar in "Bali" and sky in "Blenheim" become much worse. The SD approach suffer from similar problems. Its stereo results on "Bali" and "Blenheim" become worse after iteration due to similar inconsistency problems. They also fail to recover depths of distant objects and shape details.

Comparing to CT [7], it is observed from Figure 7 that our methods outperform theirs in both defogging and stereo. Their method still employs the conventional photo-consistency term as a stereo initialization, which becomes less effective in scattering media. Their simple addition of the stereo and defogging terms cannot reconstruct distant objects and depth details of the scene. Their defogging results are also worse compared to the others since the direct inverse of scattering is sensitive to radiometric calibration.

6. Conclusion

We formulate simultaneous video defogging and stereo reconstruction as a unified energy minimization problem. Our improved photoconsistency term explicitly models the scattering effect and makes stereo matching more robust. The matting Laplacian constraint helps to preserve shape details, especially those thin and elongated structures, which are well-known challenges for conventional stereo algorithms. We further enforce the relative depth order at neighboring pixels to be consistent with their relative fog thickness. As a result, our stereo method estimates high quality depth for scenes in scattering media, and produces temporally consistent video with enhanced visibility. Experiments on the synthetic and real data demonstrate the superior performance of our method over recent algorithms on both defogging and multi-view stereo.

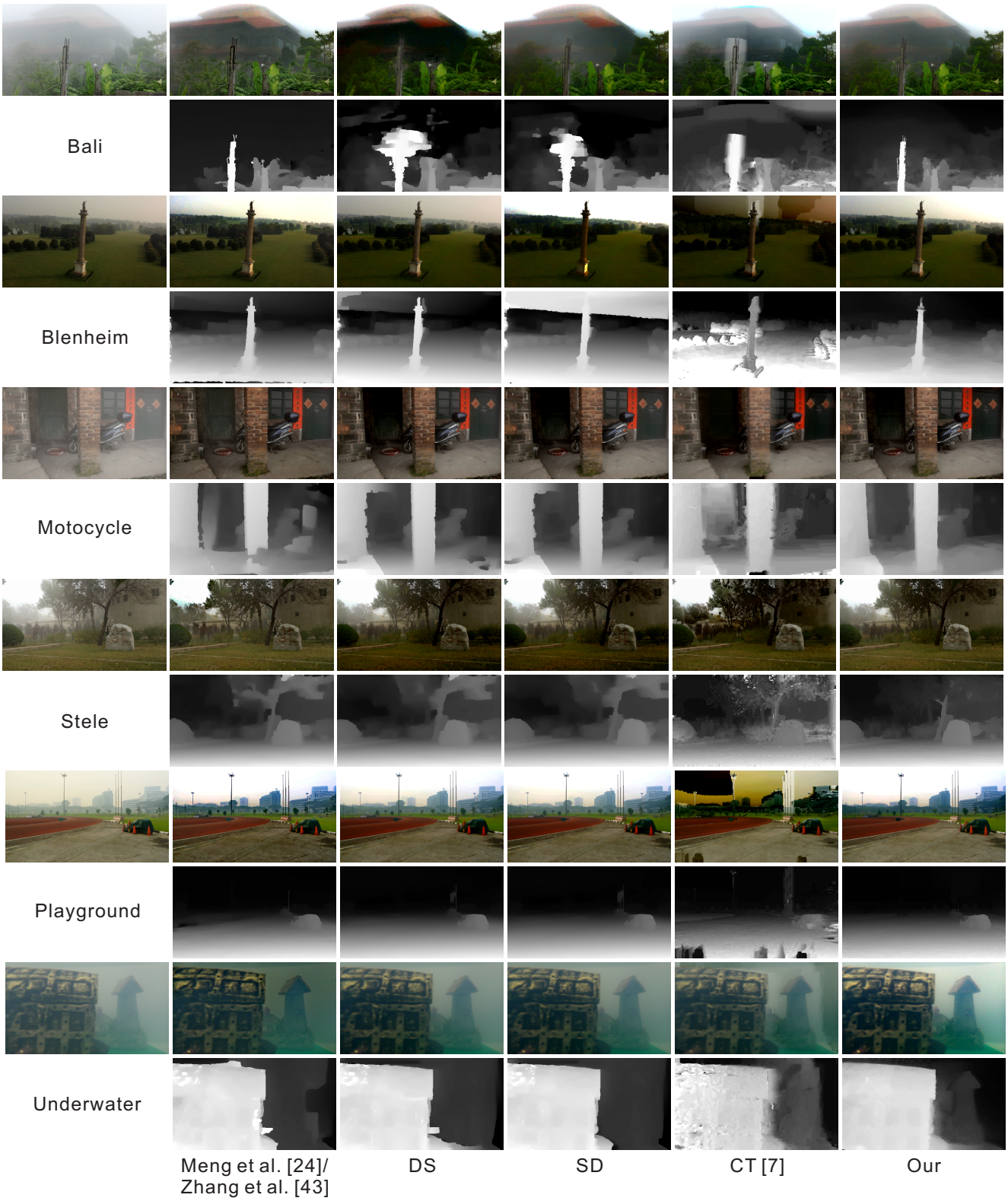

Figure 7. Comparisons of defogging and stereo results on real-world videos. First column shows the source images and the corresponding names of the data. Every odd row shows the defogging results and every even rows show the stereo results. The last row presents the name of the compared methods, corresponding to each column. Note that the second column shows the defogging result of Meng et al. [24] and stereo result of Zhang et al. [43].

Acknowledgements. This work was partially supported by R-263-000-A87-720 from Adobe Systems and Singapore PSF grant 1321202075. Ping Tan is sponsored by the Canada NSERC Discovery (No. 611664) and Discovery Acceleration Supplement (No. 611663).

References

[1] http://vision.middlebury.edu/mview/. 2

[2] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert. Enhancing underwater images and videos by fusion. In *Proc. CVPR*, 2012. 2

[3] A. F. Bobick and S. S. Intille. Large occlusion stereo. *International Journal of Computer Vision*, 33(3):181–200, 1999. 3

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, 2001. 2, 3, 5

[5] D. Bradley, T. Boubekeur, and W. Heidrich. Accurate multi-view reconstruction using robust binocular stereo and surface meshing. In *Proc. CVPR*, 2008. 2

[6] N. D. F. Campbell, G. Vogiatzis, C. Hernandez, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *Proc. ECCV*, 2008. 2

[7] L. Caraffa and J.-P. Tarel. Stereo reconstruction and contrast restoration in daytime fog. In *Proc. ACCV*. 2012. 2, 3, 6, 7

[8] P. Carr and R. Hartley. Improved single image dehazing using geometry. In *Proc. DICTA*, pages 103–110, 2009. 5

[9] F. G. Cozman and E. Krotkov. Depth from scattering. In *Proc. CVPR*, 1997. 3

[10] R. Fattal. Single image dehazing. *ACM Trans. Graph. (Proc. of SIGGRAPH)*, 27(3), 2008. 1, 2, 3, 6

[11] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *International Journal of Computer Vision*, 70(1):41–54, 2006. 3, 5

[12] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(8):1362–1376, 2009. 2

[13] D. Gallup, J.-M. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *Proc. CVPR*, 2010. 2

[14] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(3):367–383, 1992. 5

[15] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-view stereo for community photo collections. In *Proc. ICCV*, 2007. 2

[16] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. 1

[17] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(12):2341–2353, 2011. 1, 2, 3, 4, 6, 7

[18] N. Jiang, Z. Cui, and P. Tan. A global linear method for camera pose registration. In *Proc. ICCV*, 2013. 3, 6

[19] J. Kopf, B. Neubert, B. Chen, M. Cohen, D. Cohen-Or, O. Deussen, M. Uyttendaele, and D. Lischinski. Deep photo: Model-based photograph enhancement and viewing. *ACM Trans. Graph. (Proc. of SIGGRAPH Asia)*, 27(5), 2008. 1, 3

[20] D. Krishnan and R. Fergus. Fast image deconvolution using hyper-laplacian priors. In *Proc. NIPS*, 2009. 5

[21] A. Levin, D. Lischinski, and Y. Weiss. A closed-form solution to natural image matting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):228–242, 2008. 1, 2, 3, 4

[22] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 6

[23] C. Mei, G. Sibley, M. Cummins, P. Newman, and I. Reid. A constant-time effient stereo slam system. In *Proc. BMVC*, 2009. 1

[24] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan. Efficient image dehazing with boundary constraint and contextual regularization. In *Proc. ICCV*, 2013. 2, 6, 7, 8

[25] S. G. Narasimhan and S. K. Nayar. Contrast restoration of weather degraded images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(6):713–724, 2003. 2, 3

[26] S. G. Narasimhan, S. K. Nayar, B. Sun, and S. J. Koppal. Structured light in scattering media. In *Proc. ICCV*, pages 420–427, 2005. 2

[27] E. Nascimento, M. Campos, and W. Barros. Stereo based structure recovery of underwater scenes from automatically restored images. In *Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on*, 2009. 2, 6

[28] K. Nishino, L. Kratz, and S. Lombardi. Bayesian defogging. *International Journal of Computer Vision*, 98(3):263–278, 2012. 2

[29] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. In *Proc. CVPR*, 2004. 1

[30] M. Roser, M. Dunbabin, and A. Geiger. Simultaneous underwater visibility assessment, enhancement and improved stereo. In *International Conference on Robotics and Automation (ICRA)*, 2014. 2, 6

[31] Y. Y. Schechner, S. G. Narasimhan, and S. K. Nayar. Instant dehazing of images using polarization. In *Proc. CVPR*, 2001. 2

[32] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. CVPR*, 2006. 1, 2

[33] C. Strecha, R. Fransens, and L. J. V. Gool. Wide-baseline stereo from multiple views: A probabilistic account. In *Proc. CVPR*. 3

[34] C. Strecha, R. Fransens, and L. Van Gool. Combined depth and outlier estimation in multi-view stereo. In *Proc. CVPR*, 2006. 2

[35] R. Tan. Visibility in bad weather from a single image. In *Proc. CVPR*, 2008. 1, 2, 3, 6

[36] J.-P. Tarel and N. Hautiere. Fast visibility restoration from a single color or gray level image. In *Proc. CVPR*, 2009. 2, 6

[37] J.-P. Tarel, N. Hautiere, A. Cord, D. Gruyer, and H. Halmaoui. Improved visibility of road scene images under heterogeneous fog. In *IEEE Intelligent Vehicles Symposium (IV)*, 2010. 6, 7

[38] T. Treibitz and Y. Y. Schechner. Active polarization descattering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(3):385–399, 2009. 2

[39] Y. Wang, J. Yang, W. Yin, and Y. Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM J. Imaging Sciences*, 1(3):248–272, 2008. 5

[40] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. Multicore bundle adjustment. In *Proc. CVPR*, 2011. 3

[41] L. Xu, C. Lu, Y. Xu, and J. Jia. Image smoothing via L_0 gradient minimization. *ACM Trans. Graph.*, 30(6):174, 2011. 5

[42] C. Zhang, Z. Li, R. Cai, H. Chao, and Y. Rui. As-rigid-as-possible stereo under second order smoothness priors. In *Proc. ECCV*, 2014. 1

[43] G. Zhang, J. Jia, T.-T. Wong, and H. Bao. Consistent depth maps recovery from a video sequence. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(6):974–988, 2009. 2, 3, 5, 6, 7, 8