Estimating Fog Parameters from an Image Sequence using Non-linear Optimisation

Yining Ding Andrew M. Wallace
Edinburgh Centre for Robotics
Heriot-Watt University, Edinburgh, UK
{yd2007, a.m.wallace}@hw.ac.uk

Sen Wang
Sense Robotics Lab
Imperial College London, London, UK
sen.wang@imperial.ac.uk

Abstract

Given a sequence of images taken in foggy weather, we seek to estimate the atmospheric light and the scattering coefficient. These are key parameters to characterise the nature of the fog, to reconstruct a clear image (defogging), and to infer scene depth. Existing methods adopt a sequential estimation strategy which is prone to error propagation. In sharp contrast, we take a more systematic approach and jointly estimate these parameters by solving a unified non-linear optimisation problem. Experimental results show that the proposed method is superior to existing ones in terms of both estimation accuracy and precision. Our method further demonstrates how image defogging and depth estimation can be linked to a visual localisation system, contributing to more comprehensive and robust perception in fog.

1. Introduction

Safe operation is of paramount importance for modern autonomous vehicles and mobile robots. Challenges arise not only in favourable weather, but are increased significantly under adverse conditions such as fog.

Fog is caused by suspended small water droplets in air. Their interactions with light can be well explained by the atmospheric scattering model summarised in [20]. This states that the observed intensity in fog, $I(\mathbf{x})$ ($\mathbf{x}$ denotes a pixel location), is a convex combination (controlled by the transmission coefficient $t(\mathbf{x}) \in [0,1]$) of the latent clear intensity $J(\mathbf{x})$ and the atmospheric light A (Eq. (1)). $t(\mathbf{x})$ is determined by the scattering coefficient β and the distance $d(\mathbf{x})$ between a scene point and the camera (Eq. (2)). $d(\mathbf{x})$ can be related to the depth $z(\mathbf{x})$ given the camera intrinsic parameters.

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

$$t(\mathbf{x}) = \exp(-\beta d(\mathbf{x})) \tag{2}$$

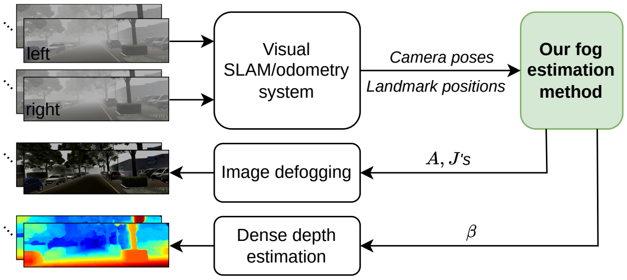


Figure 1. Our method can be used to join other systems together for more comprehensive and robust perception in fog.

Eq. (1) and Eq. (2) assume the fog to be homogeneous and therefore β and A are global values regardless of the pixel location $\mathbf{x}$. Furthermore, β is often considered to be the same for all colour channels within the visible spectrum. For simplicity, in the rest of the paper we will omit the pixel location $\mathbf{x}$ notation.

The fog parameters consist of A and β. Estimating A is a key step in almost all image defogging methods including [2, 14]. β is also an essential parameter because, as Eq. (2) suggests, it governs the mapping between the scene depth z and the transmission coefficient t. Consequently, an accurate estimate of β plays a crucial part in simultaneous defogging and stereo reconstruction methods [6, 11, 15].

In this paper, we present an elegant yet effective method which jointly estimates A and β by solving a unified non-linear optimisation problem. As by-products, our method also generates defogged pixel intensities of relevant landmarks in the map. As Fig. 1 depicts, it can be used to provide a connecting link from a visual Simultaneous Localisation and Mapping (SLAM) or odometry system to an image defogging and depth estimation system leading to more comprehensive and robust perception in fog.

Our contributions are threefold:
1) We propose a novel approach to fog parameter estimation which, unlike existing methods, jointly estimates the fog parameters by non-linear optimisation. It also requires less assumptions and achieves better performance.

2) Being versatile, our method can be plugged flexibly into most existing visual SLAM/odometry systems as an add-on module for foggy weather.
3) We demonstrate the use of our method for downstream image defogging and dense depth estimation tasks.

2. Related Work

We review existing methods of estimating fog parameters. Some early approaches estimate A from multiple images of the same scene acquired under different conditions, such as the visibility [19] or manually changed polarisation [22]. This already makes such methods inherently unsuitable for autonomous vehicle or mobile robot applications. In addition, some methods require further assumptions, such as the presence of a sky region in the images. In the rest of this section, we focus on existing work that processes a single image, a stereo pair of images or a sequence of images that an onboard camera can acquire.

2.1. Estimating A

Estimating A is critical in conventional single image defogging methods. To this end, [23] obtains A from the pixels that have the highest intensity in the input image, [14] relies on the dark channel prior to first locating the most haze-opaque region in an image then computes A from these pixel intensities, [8] estimates A as the brightest pixel value among all local minima, and [2] locates A in RGB space by leveraging the observation that fog transforms the pixel intensity distribution from tight clusters to stretched lines (dubbed 'haze-lines'). Given the limited amount of information embedded in a single image, some of these approaches have demonstrated their general effectiveness in estimating A and therefore are adopted by later conventional methods [7], and some pioneering deep learning based methods such as [4]. Even some video defogging methods such as [15] directly follow [14]'s approach in estimating A, due to its robustness and favouring its simplicity. Similarly, [5] applies firstly [8]'s method to compute an A value from the current frame. To impose temporal consistency, they then refine their estimate of A by calculating a weighted average of this A value and the A estimate from the previous frame.

2.2. Estimating β

As Eq. (1) suggests, estimating β is typically of no interest to methods that are developed for the sole purpose of defogging as long as t can be directly inferred from I (*e.g*. [2, 14]). This topic can be categorised into *perceptual* estimation and *qualitative* estimation. Methods including [9, 12, 16] achieve referenceless prediction of perceptual fog density from a single image. Although their predicted perceptual fog density indices may correlate well with human judgements, the authors make no attempt to show how these perceptual indices can be mapped to a *numerical* value of β. To our best knowledge, the *quantitative* estimation of β is little addressed in the existing literature. As Eq. (2) implies, β is the key linkage between the problem of defogging and the problem of scene depth estimation, and consequently an accurate estimate of its value plays a crucial part in various existing simultaneous defogging and stereo reconstruction methods [6, 11, 15]. In general, estimating β entails observing the same object (more precisely - the same J) from a range of known distances, which makes this task extremely challenging at best and not always possible when only a single image or even only a stereo pair of images is available. As a special case, [13] achieves β estimation from just a single image but requires the image to contain both the sky and the road, and that the road surface is homogeneous and flat (so that a known depth can be associated with each image row from the road after calibration). These are indeed very strong and application-specific constraints, which make the method impossible to be applied to general scenes. In contrast, [15] uses a sequence of images and performs structure-from-motion to facilitate observations of the same object from a range of known distances. After A is estimated following [14], they use each pair of observations whose inverse depth difference is large enough to compute a β estimate by inverting the atmospheric scattering model. Then all the estimates are gathered, from which they build a histogram of β and choose the value from the highest bin.

To summarise, using a sequence of images [5, 15] to estimate the fog parameters is much more robust compared with using a single image or a stereo pair of images [2, 8, 13, 14, 23], because more information is available and less assumptions or constraints have to be made. Nevertheless, existing methods still have a few shortcomings. [5] estimates A only and intends to assure its temporal consistency by introducing a weighted average scheme. However, as the key factor in controlling such consistency, the weight itself becomes a learnable parameter and requires fine-tuning for overall optimal performance in different scenarios. As will be shown in Sec. 4, the A and β estimation strategy proposed in [15] has the following major drawbacks. First, A is still estimated from a single image (*i.e.* the current frame), which does not make use of temporal consistency. Second, estimating β requires A to be estimated beforehand and in this way any error in A's estimate propagates to β's.

Distinct from the existing methods which estimate the fog parameters sequentially, we propose an optimisation-based method which jointly estimates A and β. Our method assumes only local homogeneity of the fog, which is acceptable in practice and widely assumed by existing methods. Unlike [13], we do not rely on any strong assumption about image content.

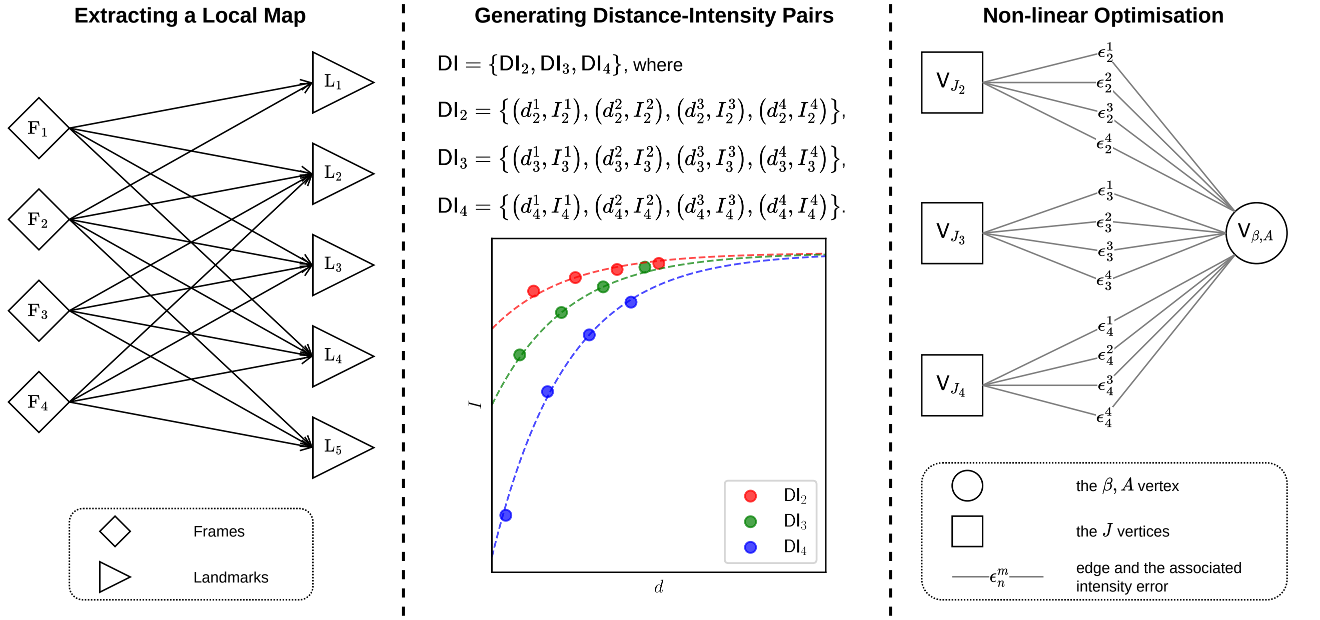

Extracting a Local Map
F_1
F_2
F_3
F_4
L_1
L_2
L_3
L_4
L_5
Frames
Landmarks
Generating Distance-Intensity Pairs
$\mathsf{DI} = \{\mathsf{DI}_2, \mathsf{DI}_3, \mathsf{DI}_4\}$, where
$\mathsf{DI}_2 = \{(d_2^1, I_2^1), (d_2^2, I_2^2), (d_2^3, I_2^3), (d_2^4, I_2^4)\}$,
$\mathsf{DI}_3 = \{(d_3^1, I_3^1), (d_3^2, I_3^2), (d_3^3, I_3^3), (d_3^4, I_3^4)\}$,
$\mathsf{DI}_4 = \{(d_4^1, I_4^1), (d_4^2, I_4^2), (d_4^3, I_4^3), (d_4^4, I_4^4)\}$.
I
d
DI_2
DI_3
DI_4
Non-linear Optimisation
V_{J_2}
V_{J_3}
V_{J_4}
$\mathsf{V}_{\beta,A}$
ϵ_2^1 ϵ_2^2 ϵ_2^3 ϵ_2^4
ϵ_3^1 ϵ_3^2 ϵ_3^3 ϵ_3^4
ϵ_4^1 ϵ_4^2 ϵ_4^3 ϵ_4^4
the β, A vertex
the J vertices
ϵ_n^m
edge and the associated intensity error

Figure 2. An overview of our three-step method. *Left* - An example of a local map containing four frames and five landmarks connected by arcs indicating observation relations (Sec. 3.1). *Middle* - The corresponding distance-intensity pairs (Sec. 3.2) and their scatter plot. DI from the same landmark share the same colour. The dashed curves are generated by Eq. (6) using groundtruth values. There is no DI associated with L_1 or L_5 because too few frames (less than four) observe them. *Right* - The corresponding optimisation problem depicted by a graph (Sec. 3.3). Note these figures are for *illustrative purposes only*. In reality, the local map typically contains many more landmarks, with each landmark observed by many more frames, and therefore the optimisation graph consists of many more vertices and edges.

3. Method

In a nutshell, our method estimates the fog parameters (*i.e.* β, A and J's) of a physical model (*i.e.* the atmospheric scattering model described by Eq. (1) and Eq. (2)) given observations (*i.e.* d and I). Fig. 2 depicts the following three steps from left to right: extracting a local map (Sec. 3.1), generating distance-intensity pairs (Sec. 3.2), and non-linear optimisation (Sec. 3.3).

3.1. Extracting a Local Map

We first build a local 3D feature map from a sequence of image frames, then generate observations in the subsequent step. The reasons why we use a local map, as opposed to a global one, are twofold: a) The optimisation problem is confined to an appropriate scale by limiting the number of observations so that it can be efficiently solved; b) Doing so implicitly assumes the fog to be only locally homogeneous, which imposes a less strong (thus more realistic) constraint.

A local map is essentially a collection of observations that sufficiently describe which local frames observe which local landmarks. It can be depicted as a directed graph, in which there exists an arc (*i.e.* a directed edge) pointing from the mth frame F_m to the nth 3D landmark L_n if and only if F_m can observe L_n. We use $\langle m, n\rangle$ to denote such an arc. Then a local map is represented as a set M, a collection of all such arcs. See the left of Fig. 2 for an example of a local map containing four frames and five landmarks.

Furthermore, we use $\mathsf{M}_n \subseteq \mathsf{M}$ to denote all inward arcs pointing towards L_n and therefore $|\mathsf{M}_n|$ is the number of frames that can observe L_n. M can be represented as a collection of such non-overlapping subsets:

$$\mathsf{M} = \{\mathsf{M}_n \mid n = 1, 2, \ldots, N\}, \tag{3}$$

where N is the total number of landmarks in the local map.

3.2. Generating Distance-Intensity Pairs

This step (see the middle of Fig. 2 for an illustrative example) prepares valid observations for the subsequent optimisation step in the form of distance-intensity pairs. We denote such observations as a set DI that is composed of a number of non-overlapping subsets DI_n:

$$\mathsf{DI} = \{\mathsf{DI}_n \mid n = 1, 2, \ldots, N \;\wedge\; |\mathsf{M}_n| \geq 4\}. \tag{4}$$

The number of such subsets is typically smaller than N because for each landmark L_n to generate DI_n, we require that at least four frames observe it. Each DI_n is a collection of distance-intensity pairs of L_n:

$$\mathsf{DI}_n = \{(d_n^m, I_n^m) \mid \langle m, n\rangle \in \mathsf{M}_n\} \subseteq \mathsf{DI}, \tag{5}$$

where d_n^m denotes the Euclidean distance between L_n and F_m, and I_n^m denotes the intensity of L_n's corresponding 2D feature point in F_m[1] (a bilinear interpolation is performed

[1] We have tested various methods of producing I_n^m (see our supplementary material) and the one described here has the best performance.

in case of a non-integer pixel location). Both d_n^m and I_n^m are typically available from a sparse feature based visual SLAM/odometry system.

3.3. Non-linear Optimisation

This step estimates fog parameters β and A, together with the landmarks' clear intensity J's, by minimising a cost function derived from the observations generated from the previous step. The problem of interest can be represented naturally as a graph (see the right of Fig. 2 for an example), in which vertices represent variables to optimise, and edges represent observation errors. An edge (*i.e.* error) connects vertices (*i.e.* variables) that contribute to the underlying error term. In such a graph for the problem of our interest, there are two types of vertices.

1. $\mathsf{V}_{\beta,A}$ encodes the fog parameters β and A. We consider β and A to be global in the case of homogeneous fog and therefore there can be only one such vertex in the whole graph.
2. V_{J_n} encodes L_n's clear intensity J_n. The number of occurrences of such vertex is the same as the number of non-overlapping subsets in DI.

According to the atmospheric scattering model, we can compute the predicted intensity value of L_n observed by F_m given d_n^m, β, A and J_n.

$$\begin{aligned} {}_{\text{pred}}I_n^m &= J_n \exp\left(-\beta d_n^m\right) + A\left(1 - \exp\left(-\beta d_n^m\right)\right) \\ &= \left(J_n - A\right)\exp\left(-\beta d_n^m\right) + A \end{aligned} \tag{6}$$

We define the intensity error ϵ_n^m to be the scalar difference between the observed intensity I_n^m and the corresponding predicted intensity ${}_{\text{pred}}I_n^m$.

$$\begin{aligned} \epsilon_n^m &= I_n^m - {}_{\text{pred}}I_n^m \\ &= I_n^m - \left(\left(J_n - A\right)\exp\left(-\beta d_n^m\right) + A\right) \end{aligned} \tag{7}$$

One can see that each ϵ_n^m can be considered as a function of β, A and J_n, and therefore adds an edge between $\mathsf{V}_{\beta,A}$ and V_{J_n} to the graph. We denote the set of all edges in the graph as E, which is composed of a number of non-overlapping subsets E_n (Eq. (8)) whose elements are intensity errors (Eq. (9)).

$$\mathsf{E} = \{\mathsf{E}_n \mid n = 1, 2, \ldots, N \text{ and } |\mathsf{M}_n| \geq 4\} \tag{8}$$

$$\mathsf{E}_n = \{\epsilon_n^m \mid \langle m, n\rangle \in \mathsf{M}_n\} \subseteq \mathsf{E} \tag{9}$$

We define each residual term to be a loss function $\ell : \mathbb{R} \rightarrow \mathbb{R}$ of ϵ_n^m. The total cost function is a weighted sum of all residual terms. Our aim is to find the best set of parameters that minimises this total cost:

$$\begin{aligned} &\underset{\beta,A,\mathsf{J}}{\operatorname{argmin}} \sum_{n:\, \mathsf{E}_n \in \mathsf{E}} \; \sum_{m:\, \epsilon_n^m \in \mathsf{E}_n} w_n^m \ell\left(\epsilon_n^m\right), \text{ subject to} \\ &\quad l_\beta \leq \beta \leq u_\beta,\, l_A \leq A \leq u_A,\, l_{J_n} \leq J_n \leq u_{J_n}, \end{aligned} \tag{10}$$

where $w_n^m \geq 0$ is the weight associated with $\ell\left(\epsilon_n^m\right)$, J is a set containing all relevant J's: $\mathsf{J} = \{J_n \mid \mathsf{E}_n \in \mathsf{E}\}$, and l's and u's are the parameter lower and upper bounds respectively. How l's and u's get set is detailed below.

Setting the bounds on the parameters

We set $l_\beta = 0.001$ and $u_\beta = 0.2$, which are rather conservative considering the equivalent visibility range is $[15, 3000]$ meters according to Eq. (15). It is worth mentioning that the same bounds are used when building the β histogram (see the middle and the right histograms in Fig. 3) in our modified version of [15] (see Sec. 4.2) to assure our method is not advantaged.

Next, for each relevant J_n we first determine if it is lower or higher than A by computing the slope k_n of the line going through the intensities observed at the maximum and the minimum distances: $k_n = \left(I_n^{d_{\max}} - I_n^{d_{\min}}\right) / \left(d_{\max} - d_{\min}\right)$. If k_n is strongly positive, we set $l_{J_n} = 0$ and $u_{J_n} = I_n^{d_{\min}}$, and we add $I_n^{d_{\max}}$ to the candidate set of l_A. If k_n is strongly negative, we set $l_{J_n} = I_n^{d_{\min}}$ and $u_{J_n} = 255$. If neither, we set $l_{J_n} = 0$ and $u_{J_n} = 255$.

Finally, we set l_A to be the median value of its candidate set, and $u_A = 255$. It is found that if u_A is set in a similar way to l_A, it will often be underestimated. We think this is caused by the fact that objects that are brighter than A are rare in a foggy scene.

Two-stage optimisation

We adopt a two-stage optimisation strategy. In the first stage, we choose ℓ to be the Huber loss ($\delta = 5$, which is empirically chosen and fixed across all experiments) in order to mitigate the effect of outlier observations. In the second stage, we choose ℓ to be the square loss and perform optimisation on inlier observations only. Furthermore, for each observation our system keeps a count of the number of times it has been identified as an inlier after the first stage of optimisation. This value c_n^m, denoted the inlier count of the nth landmark observed by the mth frame, is used to appropriately weight the corresponding residual term.

Setting the weight of the residuals

We list below the partial derivatives of Eq. (7).

$$\frac{\partial \epsilon_n^m}{\partial \beta} = d_n^m \left(J_n - A\right)\exp\left(-\beta d_n^m\right) \tag{11}$$

$$\frac{\partial \epsilon_n^m}{\partial A} = \exp\left(-\beta d_n^m\right) - 1 \tag{12}$$

$$\frac{\partial \epsilon_n^m}{\partial J_n} = -\exp\left(-\beta d_n^m\right) \tag{13}$$

We argue that the larger the intensity difference between a landmark's J and A, the more suitable that landmark is for estimating β. As can be seen from Eq. (11), the partial derivative of ϵ_n^m w.r.t. β is proportional to $(J_n - A)$. This suggests that when J_n is close to A this term will diminish, causing difficulties in finding the optimal β. Intuitively, when J is close to A, the range of the predicted intensity ${}_{\text{pred}}I_n^m$ when $d_n^m \geq 0$ flattens out and therefore ϵ_n^m contains very little information on the inference of β.

Because β appears in both the partial derivatives of ϵ_n^m w.r.t. A and J_n (Eq. (12) and Eq. (13)), we expect a more accurate estimate of β to help find the optimal A and J_n.

In light of this, we set the weight in our first optimisation stage to be the product of the following two terms: the absolute difference between the current estimate of J_n ($\hat{J}_n$) and the current estimate of A ($\hat{A}$), and the inlier count of the corresponding observation.

$$w_n^m = |\hat{J}_n - \hat{A}| \cdot c_n^m \tag{14}$$

It can be seen that the first term is landmark-dependent while the second term is observation-dependent.

In our second optimisation stage where only inlier observations are used, we set the weight to be uniform.

Results from our ablation study (Sec. 4.5) show that compared to naively uniformly weighting all residual terms in both optimisation stages, our weighting performs better in estimating both β and A.

Initialisation

To initialise β, A and J we do the following. If there have been previous estimates, which are obtained from the last successful run of our fog estimation process, we use these values to initialise. Otherwise (*i.e.* if the fog estimation process has never successfully run before, or if J_n has never been estimated before) we use our modified version of [15] (see Sec. 4.2) to initialise β and A, and use the observed intensity from the shortest distance to initialise J_n.

4. Experiments

In this section, we introduce the datasets, baseline methods and setup of experiments before presenting results.

4.1. Datasets

For evaluation, we add simulated fog to three synthetic datasets: the Virtual KITTI 2 dataset [3] (VKITTI2), the KITTI-CARLA dataset [10] and the Driving dataset (DRIVING) [17]. They all contain sequences of left and right clear intensity images as well as the corresponding left and right groundtruth depth maps. For each clear image, we first compute a distance map from its groundtruth depth map then synthesise its foggy image by applying the atmospheric scattering model (see the top row of Fig. 5 for sample foggy images). We fix A at 0.7, 0.8 and 0.9 for VKITTI2, KITTI-CARLA and DRIVING, respectively. For each dataset, six different visibility levels at $V_{\text{met}} = \{30, 40, 50, 60, 70, 80\}$ meters are tested[2]. The corresponding groundtruth β values are calculated according to Eq. (15). Note that β encodes the fog density and is closely related to the visibility (*i.e.* the meteorological optical range [21]) V_{met} in meters.

$$V_{\text{met}} = -\ln(0.05)/\beta \tag{15}$$

4.2. Competitive Methods

To our best knowledge, there is very limited existing work on estimating both A and β. First, we report the results of Berman *et al.* [2] which estimates A only. We further compare our method with the fog estimation strategy proposed by Li *et al.* [15] (estimating both A and β) as well as *our modified version* of it which will prove to be a much stronger baseline compared to the original one. The modifications we make are *twofold*. First, as proposed by [24], we use the median, instead of the maximum [14], of the 0.1% pixels with the largest dark channel values to estimate A. Second, we discard β values that are not within the range $[0.001, 0.2]$ when building its histogram. This is motivated by the observations, typically at a lower visibility, that a proportion of β values are negative and that there is a big cluster of β centred at the value of zero. In many cases the zero bin has the highest counts and therefore an erroneous β estimate would be made. Fig. 3 shows examples of unbounded (left) and bounded (middle) β histograms at 30 m visibility. As we will show later, this modified version greatly improves the original one's performance in terms of both A and β and therefore we use it as a stronger baseline.

4.3. Setup

For all comparisons, we use stereo ORB-SLAM2 [18] to facilitate multiple observations of the same landmark from a range of known distances. These observations are obtained from ORB-SLAM2's local key frames and local map points after its Local BA. The fog parameters are updated after ORB-SLAM2's local mapping thread. In other words, it happens only if a new key frame is generated. A normal frame, in contrast, does not invoke the update. See our supplementary material for implementation details including pseudo code.

Because ORB-SLAM2 is multi-threaded, inherently there is some randomness in its generation of the local key frames and the local map points. We therefore run each experiment five times and keep the median result. It is worth mentioning that although the aforementioned randomness per run still exists, we evaluate all methods in parallel in the same run to always ensure a fair comparison between them. We use the Ceres Solver [1] and choose the Levenberg–Marquardt algorithm to solve Eq. (10).

[2] See our supplementary material for a summary of the data we use for evaluation.

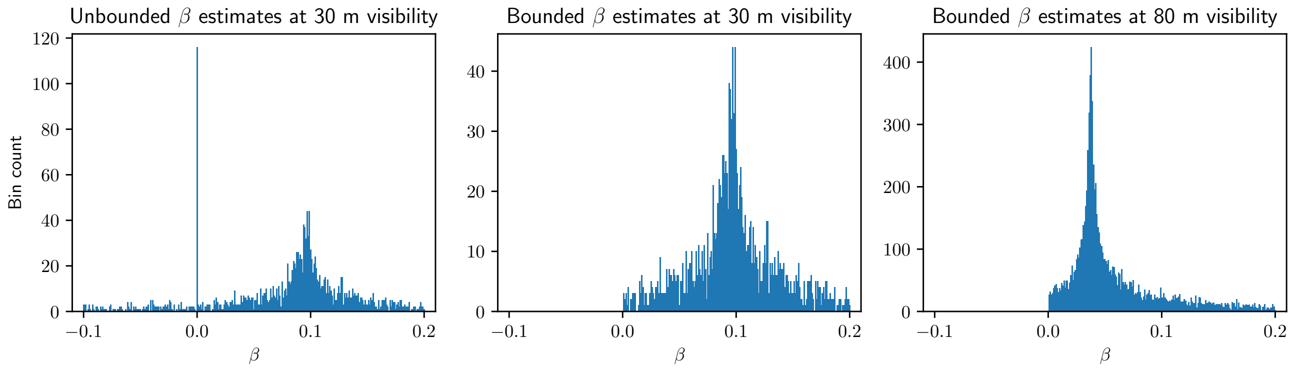


Figure 3. β histogram examples generated by Li's [15] and our modified version of it. Note that the vertical axes have different scales. *Left* - Unbounded β estimates at 30 m visibility. The highest bin which occurs at zero would lead to an erroneous β estimate. *Middle* - Bounded (within the range $[0.001, 0.2]$) β estimates at 30 m visibility. The highest bin occurs at 0.097, which is much closer to the groundtruth β value of 0.100. *Right* - Bounded (within the range $[0.001, 0.2]$) β estimates at 80 m visibility. Comparing the right to the middle, we can see that the total number of β estimates that are used to build the histogram is typically much larger at a higher visibility.

4.4. Results

We report the root-mean-square error (RMSE), the mean-absolute error (MAE) and the standard deviation (SD), in both absolute scale and relative scale, of the β and A estimates.

The top, middle and bottom of Tab. 1 show the quantitative results of the averaged β and A error metrics on VKITTI2, KITTI-CARLA[3] and DRIVING, respectively. Relative values are shown as percentages in parentheses. The results demonstrate that in most cases our method performs the best, in terms of both estimation accuracy (*i.e.* smaller errors) and precision (*i.e.* a lower standard deviation). The *only* exception is VKITTI2's A error metrics (although our β metrics are still the best in this case). A closer look at VKITTI2's results shows that our bad estimates of A stem from countryside scenes with sparse features (*e.g.* Scene02), or when the ego-vehicle is surrounded by other vehicles moving at similar speeds (*e.g.* Scene18). In either case the ORB-SLAM2's performance has been significantly degraded and therefore produces unreliable distance and/or intensity information.

Fig. 4 illustrates how β and A estimates vary with frames at various visibilities on Scene01 in VKITTI2. Each figure is generated from the run with the median RMSE of β for Li's, Li's modified and ours, and of A for Berman's. Groundtruth values are indicated by black dotted lines.

In addition, we use the open source implementation of our previous work [11] as an example to demonstrate how the fog parameters estimated by the proposed method can be used for downstream image defogging and depth estimation tasks. In [11], we use the groundtruth β in the Foggy Stereo Matching module and then estimate A in the Defogging module. We now replace both values with the ones estimated by the proposed method. Some representative results are shown in Fig. 5.

4.5. Additional Experiments

Error metrics given the groundtruth A

We test β estimation performance on KITTI-CARLA given the groundtruth A value. The quantitative results are shown in the middle block of rows in the middle subtable of Tab. 1.

We observe: a) As expected, all methods perform better when groundtruth A is given; b) For Li's and Li's modified methods, there is a significant improvement in β's error metrics when the groundtruth A is given. This is not surprising due to its *sequential* estimation strategy, since an error-free A will indeed benefit the subsequent β estimation. This observation adds to the evidence that in their method any error in A's estimate can propagate to β's; c) For our method, such improvement is much less significant. This may suggest that our method, when jointly optimising β and A with minimum prior knowledge, is able to find a β value that is not far from its best possible solution.

Ablation study

We conduct an ablation study on KITTI-CARLA to better understand how our optimisation setup affects the fog parameter estimation performance. The following three additional settings are experimented: 1) *Loose bounds*: The intensity bounds of the A and all relevant J's are trivially set to $[0, 255]$; 2) *One-stage*: Only the first stage of our optimisation is preserved, and hence all observations are treated as inliers; 3) *Uniform weight*: The weight w_n^m is fixed at 1 in both optimisation stages. The quantitative results are shown in the bottom block of rows in the middle subtable of Tab. 1.

We observe: a) Although using loose bounds marginally improves β's error metrics, A's error metrics witness a con-

[3] Result of Town04 at 30 m visibility is excluded as the ORB-SLAM2 loses tracking and provides no valid observation.

Dataset	Method	β			A		
		RMSE (%)	MAE (%)	SD (%)	RMSE (%)	MAE (%)	SD (%)
VKITTI2	Berman's	N/A	N/A	N/A	4.6808 (2.62)	3.4868 (1.95)	3.0897 (1.73)
	Li's	0.0444 (68.97)	0.0359 (54.36)	0.0247 (40.96)	5.1517 (2.89)	2.0824 (1.17)	4.7422 (2.66)
	Li's modified	0.0161 (24.20)	0.0102 (14.38)	0.0139 (21.30)	**1.8063 (1.01)**	**0.7991 (0.45)**	**1.5457 (0.87)**
	Ours	**0.0122 (17.55)**	**0.0085 (11.89)**	**0.0094 (13.92)**	3.7190 (2.08)	2.3162 (1.30)	3.2532 (1.82)
KITTI-CARLA	Berman's	N/A	N/A	N/A	12.7556 (6.25)	12.4054 (6.08)	2.8910 (1.42)
	Li's	0.0533 (89.58)	0.0493 (83.14)	0.0191 (32.52)	16.0363 (7.86)	14.9876 (7.35)	5.3108 (2.60)
	Li's modified	0.0179 (29.95)	0.0142 (23.93)	0.0126 (21.05)	10.2972 (5.05)	9.4018 (4.61)	3.7479 (1.84)
	Ours	**0.0125 (20.96)**	**0.0103 (17.33)**	**0.0084 (14.01)**	**3.0716 (1.51)**	**1.8521 (0.91)**	**2.7749 (1.36)**
	Li's (GT A)	0.0427 (71.01)	0.0335 (55.51)	0.0301 (51.39)	-	-	-
	Li's modified (GT A)	0.0129 (21.34)	0.0090 (15.04)	0.0116 (19.10)	-	-	-
	Ours (GT A)	0.0118 (20.02)	0.0095 (16.21)	0.0085 (14.46)	-	-	-
	Ours (loose bounds)	0.0121 (20.33)	0.0099 (16.75)	0.0082 (13.79)	5.4951 (2.69)	2.3398 (1.15)	5.2159 (2.56)
	Ours (one-stage)	0.0137 (22.88)	0.0114 (19.22)	0.0087 (14.41)	3.5857 (1.76)	2.2813 (1.12)	3.1537 (1.55)
	Ours (uniform weight)	0.0134 (22.66)	0.0114 (19.44)	0.0081 (13.38)	3.5878 (1.76)	2.3342 (1.14)	3.0386 (1.49)
DRIVING	Berman's	N/A	N/A	N/A	22.8736 (9.97)	14.0364 (6.12)	21.4792 (9.36)
	Li's	0.0461 (74.76)	0.0382 (61.45)	0.0264 (44.07)	14.2847 (6.22)	11.4720 (5.00)	9.0370 (3.94)
	Li's modified	0.0159 (25.04)	0.0111 (17.80)	0.0127 (20.12)	12.7639 (5.56)	9.4596 (4.12)	9.0555 (3.95)
	Ours	**0.0060 (10.66)**	**0.0044 (7.69)**	**0.0043 (7.74)**	**2.3435 (1.02)**	**1.6025 (0.70)**	**2.0229 (0.88)**

Table 1. Averaged β and A error metrics on VKITTI2 (top), KITTI-CARLA (middle) and DRIVING (bottom)

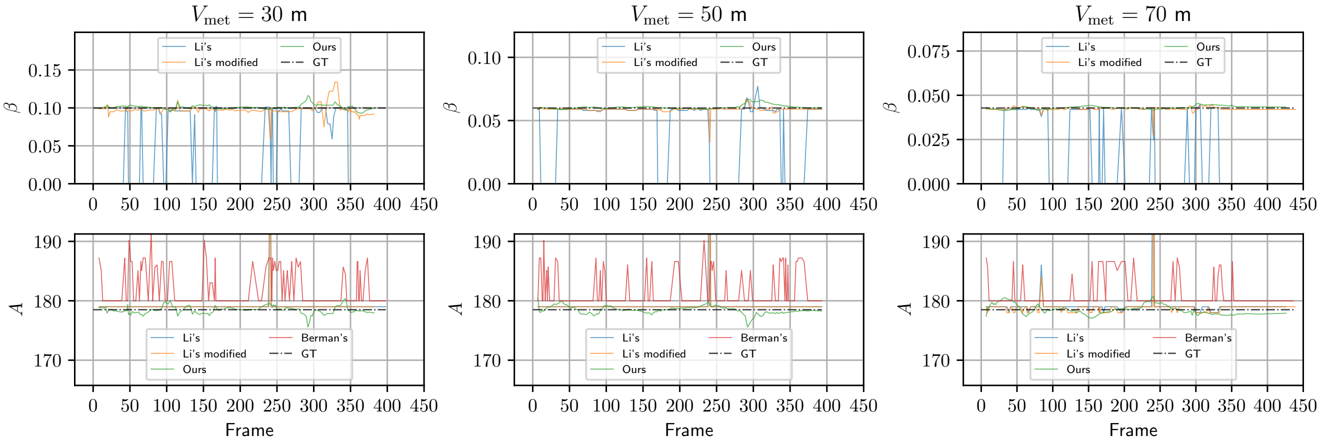


Figure 4. β and A estimates *vs*. frame evaluated on Scene01 in VKITTI2 at various visibilities (left to right: 30, 50 and 70 m). Berman's method of estimating A has many large errors. Although our modified version of Li's method significantly improves the original one's performance in both A and β estimations, both of them still have the defect that an erroneous estimate of A (*e.g*. around frame 240) would always lead to an erroneous estimate of β due to the two-step estimation strategy. See our supplementary material for more results.

siderable degradation, which happens particularly when the ORB-SLAM2 struggles to generate accurate camera poses and landmark positions; b) If one-stage optimisation is performed or a uniform weight is used, the estimation results are inferior to those produced by our full method; c) These three settings still outperform all competitive methods, despite trailing behind our full method.

Error metrics *vs*. visibility

We investigate how the fog parameter estimation performance varies with visibility. Fig. 6 plots the percentage RMSE of β (top) and of A (bottom) against visibility.

We observe: a) Our method consistently excels in both β and A estimates for all the visibilities tested; b) Both Li's and Li's modified demonstrate a downward trend in β percentage RMSE as the visibility increases. After comparing the right histogram with the middle one in Fig. 3, we infer that as the visibility grows, the number of β estimates to build a histogram becomes larger and therefore the performance of the statistics-based estimation method used by these two baseline methods improves; c) All methods witness an upward trend in A percentage RMSE as the visibility increases, which is as expected because the images will appear to be less fog-obscured as the visibility increases.

5. Conclusion

This paper presents an optimisation-based method that unifies the estimation of fog parameters in a practical setting with very few assumptions. As by-products, our method

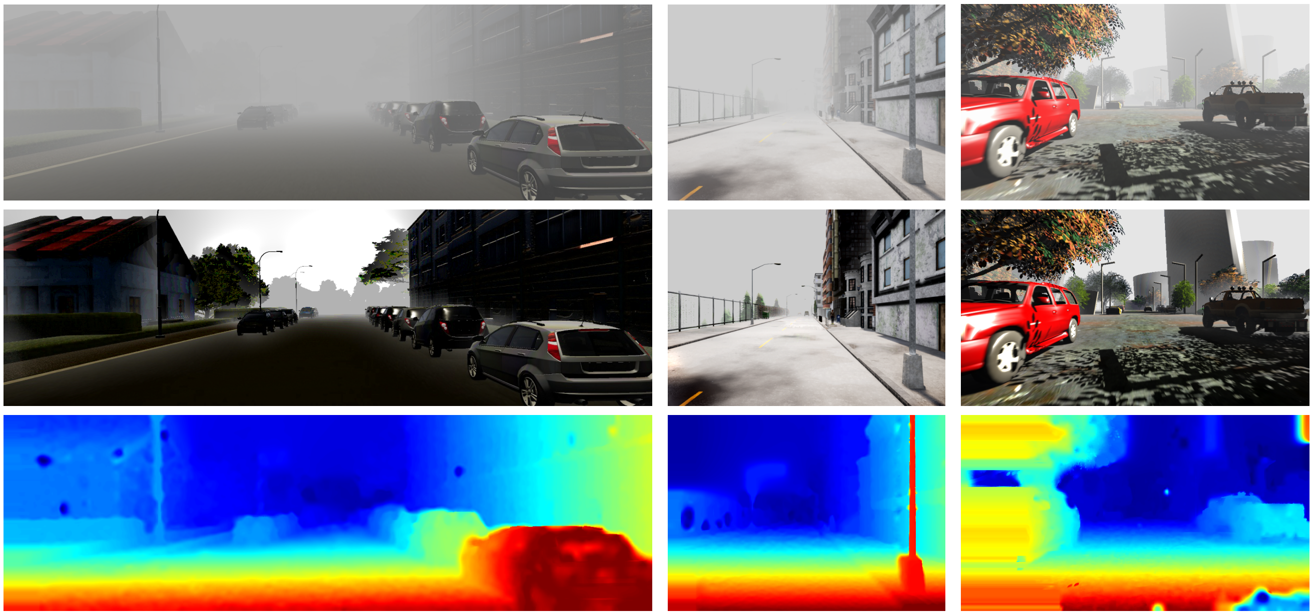
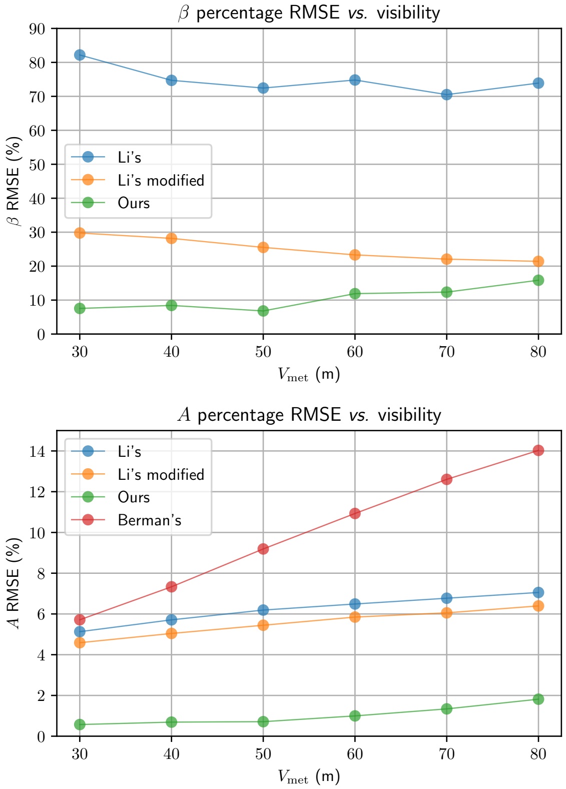


Figure 5. Top row from left to right: sample foggy images from VKITTI2 ($V_{\mathrm{met}} = 40$ m), KITTI-CARLA ($V_{\mathrm{met}} = 60$ m) and DRIVING ($V_{\mathrm{met}} = 80$ m). Bottom two rows: The defogged images and the disparity maps produced by our previous work [11] after feeding the A and β values estimated by the proposed method to it. See our supplementary material for more results.



β percentage RMSE *vs.* visibility

A percentage RMSE *vs.* visibility

Figure 6. β and A percentage RMSE *vs.* visibility on DRIVING

also defogs relevant landmarks in a map. Experimental results show that our estimation method achieves higher accuracy and precision than existing ones, which either rely on very strong constraints or are prone to error propagation. Our method can be used to provide a connecting link from a visual SLAM/odometry system to an image defogging and depth estimation system for overall more comprehensive and robust perception in fog. To illustrate this capability, we have shown representative examples produced by our previous work that makes use of the fog parameter estimation method presented here.

The distance and intensity data used by our method are obtained from a visual SLAM/odometry system. In unfavourable environments with few features and very limited visibility, such systems would typically struggle, and our fog parameter estimation performance would also be compromised. This could potentially be improved by integrating our method into a visual-inertial system that is more robust in these situations.

To the best of our knowledge, there is no suitable, existing dataset having stereo data of consecutive frames under a variety of fog densities, and with corresponding clear sequences of the same route (see our supplementary for details). Such data is necessary, both for validating estimation of the fog parameters, and for companion work in defogging and depth reconstruction. Such data collection is a priority.

Acknowledgements

We are grateful for EPSRC funding for the Centre for Doctoral Training in Robotics and Autonomous Systems EP/S023208/1.

References

[1] Sameer Agarwal, Keir Mierle, and The Ceres Solver Team. Ceres Solver, 3 2022. 5

[2] Dana Berman, Tali Treibitz, and Shai Avidan. Non-local image dehazing. In *CVPR*, 2016. 1, 2, 5

[3] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 5

[4] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE TIP*, 2016. 2

[5] Bolun Cai, Xiangmin Xu, and Dacheng Tao. Real-time video dehazing based on spatio-temporal mrf. In *Advances in Multimedia Information Processing - PCM*, 2016. 2

[6] Laurent Caraffa and Jean-Philippe Tarel. Stereo reconstruction and contrast restoration in daytime fog. In *ACCV*, 2012. 1, 2

[7] Chen Chen, Minh N. Do, and Jue Wang. Robust image and video dehazing with visual artifact suppression via gradient residual minimization. In *ECCV*, 2016. 2

[8] John Y Chiang and Ying-Ching Chen. Underwater image enhancement by wavelength compensation and dehazing. *IEEE TIP*, 2011. 2

[9] Lark Kwon Choi, Jaehee You, and Alan Conrad Bovik. Referenceless prediction of perceptual fog density and perceptual image defogging. *IEEE TIP*, 2015. 2

[10] Jean-Emmanuel Deschaud. KITTI-CARLA: a KITTI-like dataset generated by CARLA Simulator. *arXiv preprint arXiv:2109.00892*, 2021. 5

[11] Yining Ding, Andrew M. Wallace, and Sen Wang. Variational simultaneous stereo matching and defogging in low visibility. In *BMVC*, 2022. 1, 2, 6, 8

[12] Hong Guo, Xiaochun Wang, and Hongjun Li. Density estimation of fog in image based on dark channel prior. *Atmosphere*, 2022. 2

[13] Nicolas Hautiere, Jean-Philippe Tarel, Jean Lavenant, and Didier Aubert. Automatic fog detection and estimation of visibility distance through use of an onboard camera. *Machine vision and applications*, 2006. 2

[14] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE TPAMI*, 2010. 1, 2, 5

[15] Zhuwen Li, Ping Tan, Robby T. Tan, Danping Zou, Steven Zhiying Zhou, and Loong-Fah Cheong. Simultaneous video defogging and stereo reconstruction. In *CVPR*, 2015. 1, 2, 4, 5, 6

[16] Zhigang Ling, Jianwei Gong, Guoliang Fan, and Xiao Lu. Optimal transmission estimation via fog density perception for efficient single image defogging. *IEEE Transactions on Multimedia*, 2017. 2

[17] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 5

[18] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE T-RO*, 2017. 5

[19] Srinivasa G. Narasimhan and Shree K. Nayar. Chromatic framework for vision in bad weather. In *CVPR*, 2000. 2

[20] Srinivasa G. Narasimhan and Shree K. Nayar. Contrast restoration of weather degraded images. *IEEE TPAMI*, 2003. 1

[21] World Meteorological Organization. Measurement of meteorological variables, 2014. Last accessed 4 October 2022. 5

[22] Yoav Y. Schechner, Srinivasa G. Narasimhan, and Shree K. Nayar. Instant dehazing of images using polarization. In *CVPR*, 2001. 2

[23] Robby T Tan. Visibility in bad weather from a single image. In *CVPR*, 2008. 2

[24] Ketan Tang, Jianchao Yang, and Jue Wang. Investigating haze-relevant features in a learning framework for image dehazing. In *CVPR*, 2014. 5