Visual Odometry in Smoke Occluded Environments

Aditya Agarwal, Daniel Maturana, Sebastian Scherer

CMU-RI-TR-15-07

July 2014

Field Robotics Center
Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

Visual odometry is an important sensor modality for robot control and navigation in environments when no external reference system, e.g. GPS, is available. Especially micro aerial vehicles operating in cluttered indoor environments need pose updates at high rates for position control. At the same time they are only capable to carry sensors and processors with limited weight and power consumption. Thus a need arises to compare and modify state-of-the-art methods so that the appropriate one can be identified. When these MAVs are used for as tools for inspection and damage assessment, the need to navigate in challenging and degraded indoor environments becomes essential.

The work addresses the problem of odometry failure in unfavourable conditions of fire and smoke. The reasons for odometry failure are identified and various image enhancement techniques are implemented and compared. The case of contrast enhancement using image depth maps is inspected closely in particular since 3D depth data is available for use. Apart from visually enhancing the *hazy* image, the method also shows improvement in feature extraction, feature matching and inlier detection, all of which are essential components of visual odometry methods. Two visual odometry methods SVO to Fovis are then compared using various benchmarking and evaluation methods with the purpose of determining the more efficient and accurate method.

Contents

1 Introduction

1.1 Visual Odometry

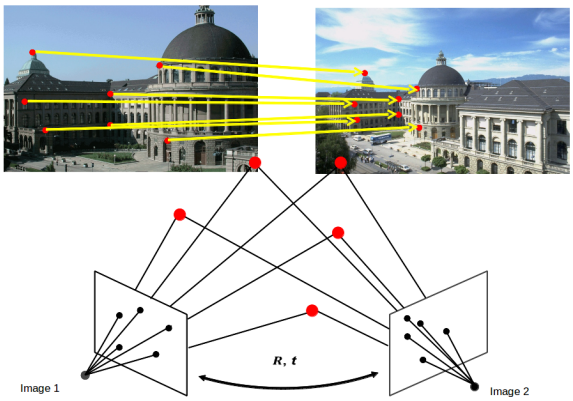


Figure 1: [13]

Micro Aerial Vehicles (MAVs) will soon play a major role in disaster management, industrial inspection and environment conservation. For such operations, navigating based on GPS information only is not sufficient. Precise fully autonomous operation requires MAVs to rely on alternative localization systems to obtain knowledge of the surrounding 3D environment for the purpose of navigation. In other words, the MAV needs to estimate its current position and orientation in the environment by relying on sensors such as cameras. RGB-D cameras capture RGB color images augmented with depth data at each pixel.

1.2 Odometry in Degraded Environments

The environment under consideration consists of passageways of a ship under unfavourable conditions of fire and smoke. This requires the construction of a small and light aerial vehicle which can navigate the narrow corridors of a ship. These conditions impose sever restrictions on visual odomoetry methods.

Construction of a suitable MAV requires limited use of hardware which means a restriction of computing capacity available. Odometry methods require significant CPU resources and hence the need arises to find a method that minimizes CPU consumption.

Introduction of smoke occluded vision introduces significant complications. Camera based odometry rely heavily on information extracted from captured sequence of images, which is contained in the form of *features*. These *features* present in consecutive image sequences are compared to estimate position and orientation of the MAV (*matching*). Introduction of smoke severely degrades image quality thus affecting the number of usable *features* and extent of *matching* that can be performed. However since RGB-D cameras illuminate a scene with an structured light pattern, they can estimate depth even in areas with poor visual texture. Thus a need arises to enhance the

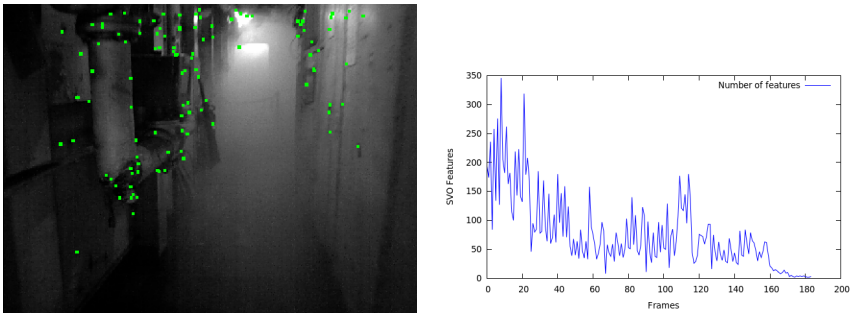
image with the objective of improving *feature extraction* and *matching*.

1.3 Image Dehazing

The process of enhancing the image occluded by smoke/fog/haze is commonly referred to as image de-hazing. Due to smoke/haze the irradiance or reflected light received by the camera gets attenuated along its line of site. Further, incoming light is reflected by suspended particles of the smoke or haze, creating a component of light called *airlight*. The degraded image loses contrast and color precision and saturation. The amount of scattering depends on the distance of the scene from the camera and hence the image degradation is spatial-variant, in other words depending on the distance. Classical contrast enhancement techniques are space-invariant, that is they act on the image as a whole, by applying the same operation for every pixel, without considering the actual distance of the scene from the observer. Thus these are not reliable solutions to the dehazing problem. Hence the need arises for dedicated image enhancement techniques. The spatial dependence also hints at utilization of depth data obtained from RGBD cameras for image enhancement.

2 Problem Statement

Prescence of smoke in the environment loss of color contrast, precision and saturation and hence visual odometry algorithms are unable to find distinguishing features in the image. In Figure 2(a) the colored dots represent detected features in the image. It can be seen from Figure 2(a) that features are identified only in haze free regions of the image. In Figure 2(b), X-axis represents the frame number in the captured video sequence while the Y-axis represents the number of features detected in the frame. The starting point of the sequence is the frame show in Figure 2(a) and as frame number increases, the camera moves closer to the haze areas. It is observed from Figure 2(b) that the number of extracted features decrease drastically as the camera moves into the haze areas.



(a) (b)

Figure 2: Variation in number of features

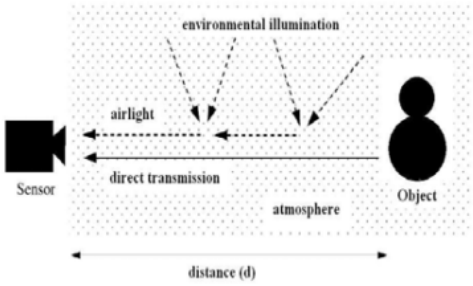


Figure 3: Illumination model in a haze environment [17]

Image dehazing can be represented physically by [3, 9, 10, 15]

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

where $I(x)$ is the input image, $J(x)$ is the scene radiance or albedo, i.e., the light reflected from its surfaces, and $x = (x, y)$ denotes the pixel coordinates. The direct transmission of the scene radiance, $t(x)J(x)$, corresponds to the light reflected by the surfaces in the scene and reaching the camera directly, without being scattered. The *airlight*, $(1 - t(x))A$, corresponds to the ambient light that causes a shift in the unscattered scene radiance. The atmospheric light vector A describes the intensity of the ambient light. For a homogeneous atmosphere, the transmission $t(x)$ can be expressed as:

$$t(x) = e^{-\beta d(x)} \tag{2}$$

where $d(x)$ is the distance from the observer.

The image dehazing problem can be thought of as recovering J.

3 Related Work

3.1 Image Dehazing

Existing image dehazing techniques can be divided into the following categories:

1. Multiple Images: The dynamic nature of smoke that allows areas of the image to be partially visible has been utilized [1] to construct a mosaic of different images of the same scene. The image is divided into regions and for each region the clearest frame is identified. The resulting image is a mosaic of all such clear frames.

2. Single Image: Intensity of haze present in an image is dependant on the distance from the observer. Hence haze removal is an underconstrained problem if only a single image is given without any depth information. Thus single haze removal methods can be classified into two groups depending on whether they use rough depth information from available 3D models or not.

   (a) Model Based: Scattering models that describe the colors and contrasts of a scene under haze conditions alongside depth information [12, 11] can be used to dehaze images. Depth information however is recovered through existing models [8] or through user input[12]. Required depth information need not be precise, accuracy upto a certain scale is observed to be sufficient. The methods can be used for both color and gray-scale images.

   (b) Contrast Enhancement: These methods rely only on single images. The success of these methods lies in using a stronger prior or assumption. These approaches impose constraints on the scene albedo and treat scene depth as a by-product of the estimation process. In other words, some methods can also recover a depth map from the varying intensity of the haze in the image. Tan [16] imposes a locally constant constraint on the albedo values (the original colors in the image) to increase the contrast in local block regions of the image. Tarel and Hautiere [18] estimate the atmospheric veil," an image of the scattered *airlight*, by using combinations of min, max, and median filters to enforce piecewise constant, and use the estimate to obtain a contrast enhanced image of the scene. Fattal [3] assumes that the surface Lambertian shading factor and the scene transmission are locally independent in order to separate the haze from the scene, and then uses a Gaussian-Markov random field to smooth the transmission values.

   (c) Dark Channel : A more recent method [6] which uses statistical properties of the original image for dehazing. It is observed that haze-free images have atleast some pixels referred to as dark pixels have very low intensity in atleast one channel (rgb). However in a haze image intensity values of these pixels is due to *airlight* and this difference is used to estimate the haze transmission and also recover a depth map of the scene.

3.2 Odometry Methods

The standard odometry approach is to extract a set of salient image features in each image and match them in successive frames using invariant feature descriptors; recover both camera motion and structure using epipolar geometry; refine the pose and structure through reprojection error minimization. Other methods estimate the motion directly from image intensity values and are sometimes more accurate even in case of poor image quality. Fast semi-direct monocular visual odometry [5] combines the advantages of both methods. The monocular version uses feature extraction to select *keyframes* or frames that are significantly different and store them in a map. This is done in the *mapping thread* responsible for maintaining and creating a 3D map of the environment. At each keyframe selection, a probabilistic depth-filter for each 2D feature is initialized. The 3D depth of this feature is not known. The filters converge using a Bayesian update step in a separate computation thread and after convergence the depth estimate is stored alongside its feature. Now this 3D scene information is used by the motion estimation thread which consists of sparse model-based image alignment, feature alignment and pose and structure refinement.

Fovis [7] is a visual odometry method designed for RGBD cameras and is similar in execution to SVO up until feature extraction. However unlike Svo, after feature extraction Svo follows the conventional feature matching process to detect candidate points that can be used for motion estimation. These points are further filtered through an inlier detection step which uses geometric verification to remove the outliers. Here the fact that 3D distance between two points does not change substantially after a rigid body motion is used to determine the fact that the two points actually correspond to the same point in the image. This step is crucial as it directly determines the capibility of *fovis* to track efficiently in a given environment.

3.3 Odometry Comparison

In [14] a complete benchmark that can be used to evaluate visual SLAM and odometry systems on RGB-D data is provided. The dataset consists of sequences recorded in two different indoor environments. Each sequence contains the color and depth images, as well as the ground truth trajectory data collected using a motion capture system. The work and [2] discuss evaluation techniques and metrics that can be used to compare accuracy of visual odometry systems.

4 Approach

The problem is approached in two steps. First image dehazing methods are compared and evaluated and the most suited methods are selected to construct an image dehazing pipeline. Then the visual odometry methods are compared, first on datasets without any dehazing and then on the same datasets with dehazing methods integrated. Section 4.1 and 4.2 describe the dehazing and odometry approaches used.

4.1 Image Dehazing Pipeline

In single image dehazing without using any depth-models [16, 18] recovering J in Equation 1 is done by first estimating the transmission, $t(x)$ and the airlight vector A. The estimated transmission is further used to recover the depth map.
Rearranging Equation 1

$$J(x) = [I(x) - (1 - t(x))A]t(x)^{-1} \tag{3}$$

Narasimhan [11] replaces the airlight vector A by $I(x)_\infty$, which for the case of atmospheric hazing referes to the sky brightness.

For the purpose of this work, both $I(x)_\infty$ and A determined by varying their values against the number of SIFT features present in the corresponding dehazed image.

The dehazing methods used in the pipeline are selected after rigorous comparison between available dehazing methods (Section 5.1). As shown in Figure 4, a combination of contrast enhancement [18] and depth based image enhancement [11] is used to enhance the overall image. Since we have depth data available from the RGBD camera, [11] is an obvious choice as it makes direct use of available depth data. A median filter is used to smooth out the depth artificats in the depth data available. Contrast enhancement [18] enhances image areas for which depth data is unavailable. In place of contast enhancement, RGB based enhancement [6] can also be used if RGB input/dataset is available.

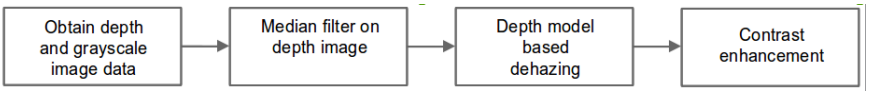


Figure 4: Image Dehazing Pipeline

4.2 Visual Odometry

The image dehazing pipeline is integrated with two representative visual odometry methods - $fovis$ [7] and svo [5]. Comparison between odometry methods is performed by using TUM RGBD Benchmarking tools [14] by calculating relative pose error which measures the local accuracy of the trajectory over a fixed time interval. It corresponds to the drift in the trajectory and the automated script evaluation available allows for easy evaluation using trajectory and pose data exported from ROS. The absolute trajectory error measures deviation from the ground truth available for TUM

datasets and is used to construct the plots. For shadwell datasets, localization data is used as ground truth.

4.2.1 Svo using RGBD

The monocular version of svo fails for cases when there's a challenging scene (e.g. motion blur, darkness, lack of keypoints). This happens at the map reprojection and feature alignment step where the code is unable to find enough matching features after reprojection. When the algorithm loses track of the position it tries to relocalize by locating the closest matching keyframe in the map to the current frame and then it can't relocalize either because in our case the algorithm isn't revisiting older places, just going new ones. Thus the algorithm has to be modified to work with RGBD data.

There are two pipelines used by svo. One is the mapping pipeline, which estimates the depth of 2D features in the image. This depth is then used in the monocular motion estimation pipeline which then estimates motion using 3D scene data. In the rgbd modified version the monocular pipeline from the code is used as it is and instead of using the mapping pipeline, depth information from the camera is provided directly. Following major modifications need to be made:

- **Removing depth filter**: Depth-filter is removed completely and instead new 3d points are initialized directly from the depth data obtained from the camera. Thus as each keyframe is added, instead of initializing new depth-filters, features are extracted and corresponding depths are stored alongside directly.
- **Modifying map initialization**: svo creates an initial 3D map of the scene by taking two keyframes and triangulating the 3D distances. This is no longer need in the RGBD version as the 3D map can be initialized directly from the first frame data itself.
- **Modifying keyframe selection**: It is based on relative euclidean distance to the previous frame. A keyframe is selected if the Euclidean distance of the new frame relative to all keyframes exceeds 12% of the average scene depth. However keyframe selection should also be based on how much the frame has rotated relative to the prior one. So if the frame has moved (euclidean) or rotated a lot, there should be a new keyframe.

5 Experimental Results

In order to find a suitable image dehazing and visual odometry approach, we first begin by comparing image dehazing methods. Dehazed datasets are constructed using different methods, which are then further used for comparison of odometry methods. Odometry methods are also analyzed using haze-free datasets to evaluate and compare their perfomance in an ideal setting.

5.1 Image Dehazing Comparison

- **Multiple Image Based [Donate2006] [1]** This method requires multiple views of the same smoke occluded scene. Different areas of the image are partially visible accross these views. A mosaic is constructed by combining different views, the images to be combined to enhance a certain area are selected by using color saturation and high frequency content as a metric to determine smoke occluded image quality.
- **Contrast Enhancement Based [Fattal2014] [4]** This method uses the color lines property of an RGB image to enhance contrast. Color lines are one dimensional distributions of pixels in RGB space. Variations in color lines are used to estimate the scene transmission.

Figure 5: Comparison of original and dehazed image [Fattal2014]

- **Contrast Enhancement Based [Tarel2009] [18]** A combination of median filters is used to enhance the overall image contrast. However the method is slow for real time application.
- **RGB Based [He2010] [6]** This method as shown in Figure 8, is effective for RGB images. It uses the fact that haze free images have dark pixels that have low intensity in atleast one channel. It is used to estimate transmission from a haze image and also recover an approximate depth map which is then used for dehazing. The method is not suitable for grayscale images and leads to darkening of overall image as seen in Figure 7

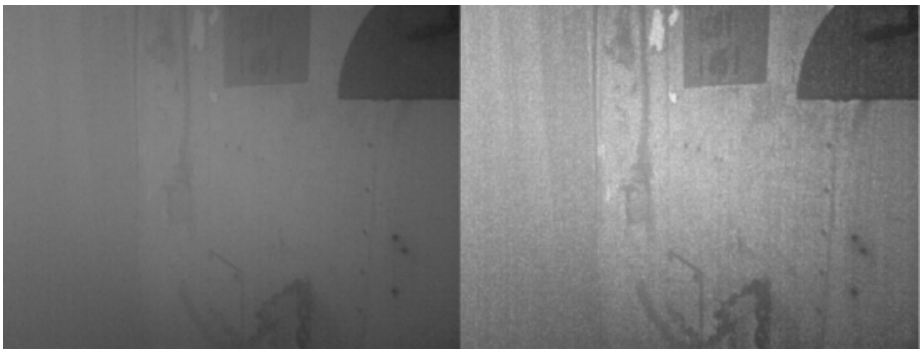


Figure 6: Comparison of original and dehazed image [Tarel2009]

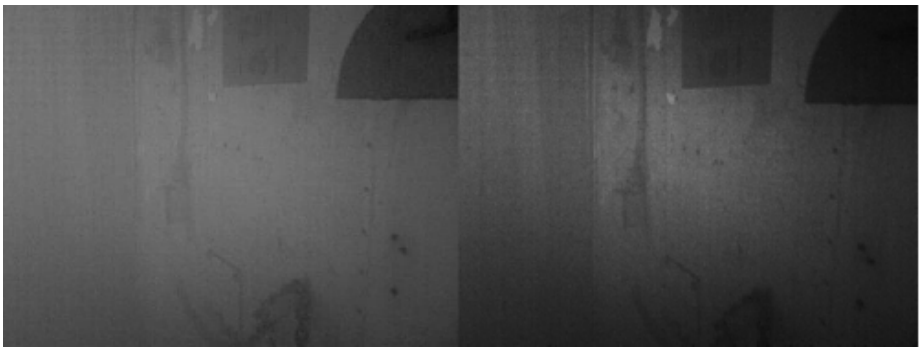


Figure 7: Comparison of original and dehazed image using RGB Based method [He2010]

- **Depth Model Based [Narasimhan2003] [11]** From Figure 9 it can be seen that significant enhancement is observed for areas for which depth information is available. Scattering coefficient β in Equation 2 is selected as 41 and I_∞ in Equation 3 is set as 77. Parameter values are selected by plotting SIFT features in the image against the parameters individually and then selecting the parameter for which maximum features are present.

- **Image Dehazing Pipeline** Figure 10 shows the improvement resulting from the application of the image dehazing piepline explained in Figure 4 on a Shadwell dataset containing haze. The second figure in the three figure image represents the image dehazed using depth model based enhancement and then median filtered to reduce artifacts. The last image represents the final image after application of contrast based enhancement.

Figure 8: Comparison of original and dehazed image using RGB Based method [He2010]

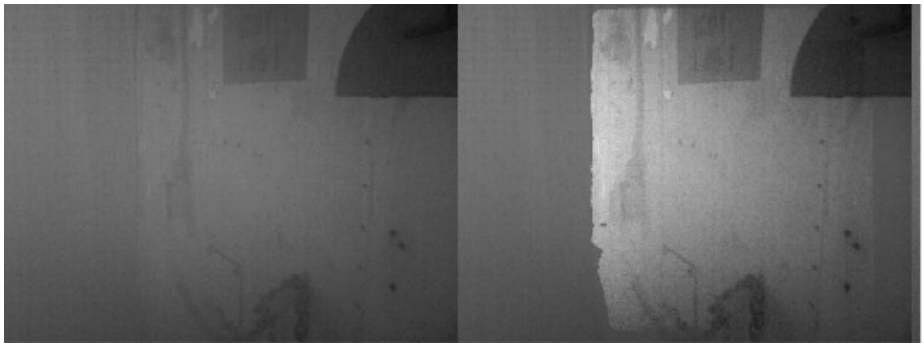


Figure 9: Comparison of original and dehazed image [Narasimhan2003]

5.2 Haze-free Odometry Comparison

The first comparison under haze-free conditions is performed using benchmarks datsets and tools presented in [14]. The dataset used $fr2/desk$ is a relatively ideal dataset with image frames containing sufficient number of features. TUM datasets are provided along with ground truth information. Table 1 show the calculated relative pose errors. The relative pose error measures local accuracy across consecutive frames relative to the ground truth. It is a root mean square error that gives an indication of drift of trajectory. Thus a higher error means more tendency to drift over time.

Columns 1 to 4 in Table 1 show the calculated relative pose errors for TUM Datasets [14]. The less magnitude of RMSE error values suggests that both $fovis$ and svo follow the ground truth trajectory closely. It can also be seen that the difference in error values of $fovis$ and svo is negligible, indicating that performance difference is minimal in conditions where sufficient image features are available. Figure 11 shows the odometry obtained from the TUM datasets plotted against the ground truth. Figure 11

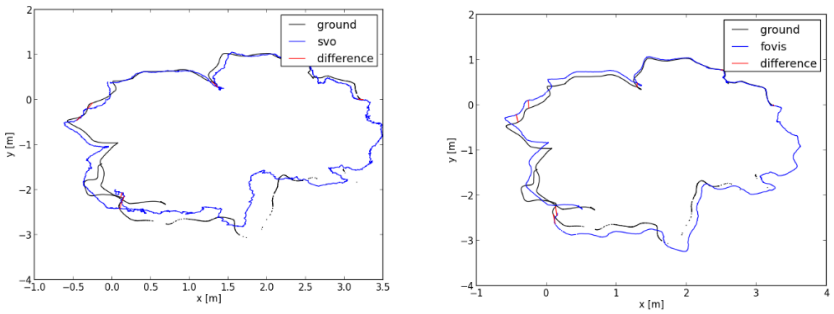
Figure 10: Dehazing using the dehazing pipeline

(a) shows odometry obtained using svo while Figure 11 (b) shows odometry obtained using $fovis$. Figure 11 (a) and (b) again establish that both $fovis$ and svo follow the ground truth trajectory closely. Columns 5 to 8 show the relative pose errors calculated using Shadwell 03_level2 dataset. The localization data has been used as ground truth for the calculation. The translational error for svo at 0.0435 is considerably more than $fovis$ at 0.0105 suggesting that $fovis$ follows the ground truth trajectory more closely, while svo has a tendency to drift over time. This is also established from Figure 12 (a) and (b) which show the trajectory plot for svo and $fovis$ respectively using the Shadwell 03_level2 dataset.

A comparison of computational efficiency showed that both svo and $fovis$ consume upto 25% of CPU or approximately an entire core of a quad-core Odroid board. However svo is signficantly faster with a runtime of $4\ ms$ as compared to $fovis$ with a runtime of $23\ ms$.

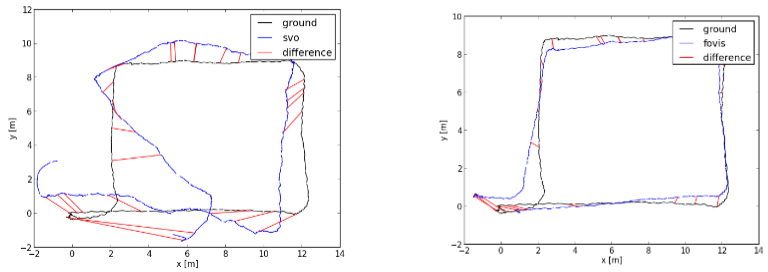
Table 1: Comparison of odometry methods under haze-free conditions

	fr2/desk				shadwell/03_level2			
Method	Translational (RMSE)	Translational (Standard dev)	Rotational (RMSE)	Rotational (Standard dev)	Translational (RMSE)	Translational (Standard dev)	Rotational (RMSE)	Rotational (Standard dev)
fovis	0.0116	0.0048	0.58	0.29	0.0178	0.0105	1.435	0.946
svo	0.0139	0.008	0.69	0.35	0.0435	0.0288	1.417	0.944



(a) (b)

Figure 11: (a)svo against ground trajectory on TUM dataset (b)$fovis$ against ground trajectory on TUM dataset



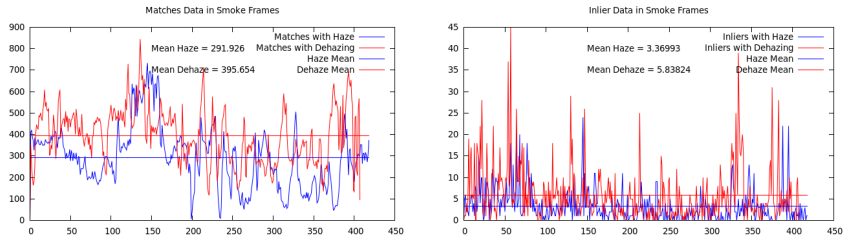
(a) (b)

Figure 12: (a)svo against ground trajectory on Shadwell dataset (b)$fovis$ against ground trajectory on Shadwell dataset

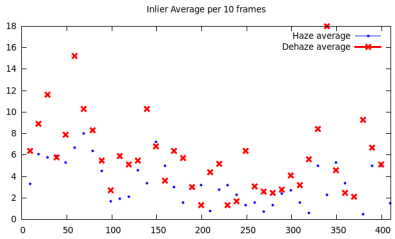
5.3 Integrated Image Dehazing and Visual Odometry

As explained before $fovis$ uses feature matching followed by inlier detection to determine candidate points that can be used for motion estimation. Thus the average number of matches and inliers detected across frames is a good indication of the performance of $fovis$. Hence we compute the average number of matches and inliers in datasets with haze and then in datasets dehazed using the image dehazing pipeline in Figure 4. The corresponding comparison is presented in Figure 13. Figure 13(a) and (b) show the number of matches and the number of inliers on the Y-axis respectively, calculated across image frames in the dataset. The red line represents dehazed dataset while blue line represents datasets with haze. It can be seen that number of matches as well as number of inliers detected has improved after when $fovis$ is run on the dehazed dataset. Figure 13 (c) is plotted by computing average number of inliers for every 10 frames. It can be seen that the red marks representing the dehazed dataset lie above the blue ones represting datasets with haze, indicating an increase in the number of inliers detected and hence an improvement in performance of $fovis$.

Feature extraction is also performed by svo and hence the number of features extracted are an indication of svo performance. Figure 14 (b) represents features extracted by svo in a dehazed dataset and the same dataset with haze, represented by red and blue lines respectively. It can be seen that the number of features extracted in a dataset with haze are negligible and the number of features extracted have significantly increased in the case of dehazed dataset. This is also established from Figure 14(a) which shows features are extracted only in dehazed regions of the image. Another indicator of svo performance is the number of reprojection matches, that is the number of matches obtained by reprojecting current frame onto the map. Figure 14(c) shows the number of reprojection matches for dehazed dataset and dataset with haze. The figure shows significant improvement in number of matches detected when the dehazed dataset is used.

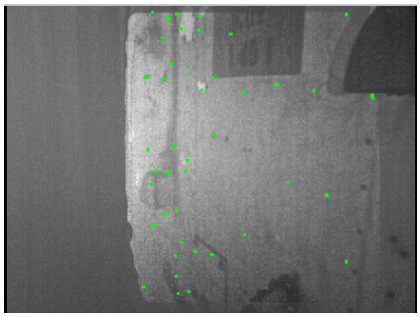


(a) (b)

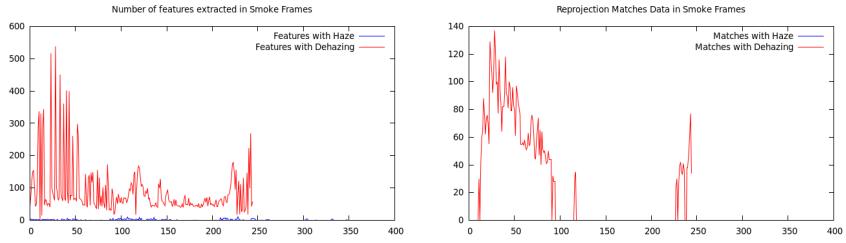


(c)

Figure 13: Improvements in $fovis$ (a) Number of matches across frames (b) Number of inliers across frames (c) Average inliers every ten frames



(a)



(b) (c)

Figure 14: Improvements in svo (a) Feature detection by svo in a dehazed image (b) Number of features across frames (c) Number of reprojection matches across frames

6 Conclusion

It can be concluded that svo and $fovis$ are both suitable odometry methods for a light MAV, $fovis$ being the more suitable candidate if more accuracy is desired and svo if speed. For smoke occluded environments a combination of depth and contrast based dehazing can enhance input image and hence significantly improve performance of semi-direct odometry methods such as svo. Improvement in feature based methods such as $fovis$ may not be as significant.

7 Future Work

The image dehazing pipleline in Figure 4 can be altered by using a colored smoke dataset and thus replacing the contrast enhancement step by RGB based enhancement. For the depth based enhancement step in the pipeline, automatic selection of parameters according to the environment can be included. Also recording of more ideal smoke datasets will enable evaluation of more dehazing methods.

References

[1] Arturo Donate and Eraldo Ribeiro. Viewing Scenes Occluded by Smoke. pages 1666–1675, 2006.

[2] Felix Endres, Jurgen Hess, Nikolas Engelhard, Jurgen Sturm, Daniel Cremers, and Wolfram Burgard. An evaluation of the RGB-D SLAM system. *2012 IEEE International Conference on Robotics and Automation*, 3(c):1691–1696, May 2012.

[3] Raanan Fattal. Single image dehazing. *ACM SIGGRAPH 2008 papers on - SIGGRAPH '08*, page 1, 2008.

[4] Raanan Fattal. Dehazing using color-lines. New York, NY, USA, 2014. ACM.

[5] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO : Fast Semi-Direct Monocular Visual Odometry. 2014.

[6] Kaiming He, Jian Sun, and Xiaoou Tang. Single Image Haze Removal Using Dark Channel Prior. *IEEE transactions on pattern analysis and machine intelligence*, August 2010.

[7] Albert S. Huang, Abraham Bachrach, Peter Henry, Michael Krainin, Dieter Fox, and Nicholas Roy. Visual odometry and mapping for autonomous flight using an rgb-d camera. In *In Proc. of the Intl. Sym. of Robot. Research*, 2011.

[8] Johannes Kopf, Boris Neubert, Billy Chen, Michael Cohen, Daniel Cohen-or, Oliver Deussen, Matt Uyttendaele, and Dani Lischinski. Deep Photo : Model-Based Photograph Enhancement and Viewing.

[9] S. G. Narasimhan and S. K. Nayar. Chromatic framework for vision in bad weather. *CVPR*, pages 598–605, 2000.

[10] S. G. Narasimhan and S. K. Nayar. Vision and the atmo- sphere. *IJCV*, 48:233254, 2002.

[11] S.G. Narasimhan and S.K. Nayar. Contrast restoration of weather degraded images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6):713–724, June 2003.

[12] Srinivasa G Narasimhan and Shree K Nayar. Interactive (De) Weathering of an Image using Physical Models . pages 1–8, 2003.

[13] Davide Scaramuzza and Friedrich Fraundorfer. Visual odometry: Part i the first 30 years and fundamentals. *IEEE Robotics and Automation Magazine*, 18(4), 2011.

[14] Jrgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 573–580, October 2012.

[15] R. Tan. Visibility in bad weather from a single image. *CVPR*, 2008.

[16] Robby T. Tan. Visibility in bad weather from a single image. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[17] RobbyT. Tan. Dehazing and defogging. In Katsushi Ikeuchi, editor, *Computer Vision*, pages 174–177. Springer US, 2014.

[18] Jean-philippe Tarel and Nicolas Hauti. Fast Visibility Restoration from a Single Color or Gray Level Image. 2009.