Robust Dense Optical Flow with Uncertainty for Monocular Pose-Graph SLAM

Yonhon Ng, Jonghyuk Kim and Hongdong Li
Research School of Engineering
The Australian National University, Australia
{yonhon.ng, jonghyuk.kim, hongdong.li}@anu.edu.au

Abstract

In this paper, we propose how to use dense optical flow field as opposed to sparse feature matches to improve large-displacement monocular visual odometry. The principled framework we developed incorporates uncertainties in the construction of a four-dimensional cost volume for dense flow computation. A novel weighted eight-point algorithm is derived which robustly estimates inter-frame camera motions by using the obtained dense correspondences with uncertainties. This initial motion estimation is subsequently employed to achieve potential loop closing operation, optimised jointly in a robust pose-graph SLAM framework. Performance of the proposed new method has been validated on standard benchmark dataset – KITTI dataset. Experimental results demonstrate that the accuracy of our method is on par with other state-of-the-art methods without relying on commonly used priors such as motion constraint or ground plane segmentation.

1 Introduction

Pose-graph based SLAM (simultaneous localisation and mapping) algorithms have drawn significant interest in both robotics and computer vision communities [Kümmerle *et al.*, 2011; Stachniss and Kretzschmar, 2017]. These algorithms typically consist of a front-end and a back-end part as illustrated in Figure 1. For a vision-based SLAM system, the front-end typically performs visual odometry between consecutive images or relative to the map, and the back-end aims to optimise the pose graph, as well as screening out outliers.

The visual odometry system can generally be categorised into monocular camera based [Geiger *et al.*, 2011; Fanani *et al.*, 2017], stereo based [Mur-Artal and Tardós, 2016; Engel *et al.*, 2015], multi-camera or omnidirectional camera based [Meilland *et al.*, 2011; Tardif *et al.*, 2008] and RGBD camera based systems [Endres *et al.*, 2012; Handa *et al.*, 2014]. Monocular visual odometry is the most challenging among them, partly due to the limited field of view and the inherent scale-indeterminacy in translation estimation [Hartley and Zisserman, 2003].

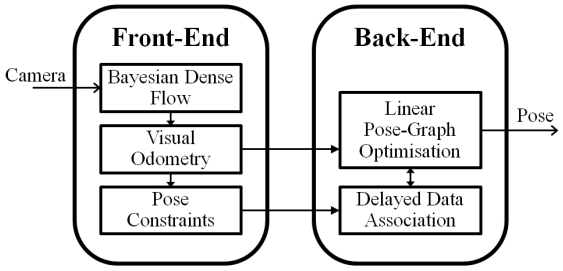

Figure 1: Framework of our proposed pose-graph SLAM

Despite this, monocular visual odometry is popularly adopted by many robotics applications (e.g. a mobile robot platform equipped with a single video camera), due to its low cost and its ease of construction and maintenance.

[Fanani *et al.*, 2017] proposed a monocular visual odometry method that does not rely on random sampling and consensus (RANSAC) algorithm or multi-frame bundle adjustment (BA) that most existing method uses. Instead, they use convolutional neural network (CNN) based ground segmentation and motion predictive model [Bradler *et al.*, 2015] to obtain a robust scale estimation result that is almost on par with stereo methods. The use of CNN for ground segmentation implies that the method may not be suitable for general use, when the robot or car travels on a surface that looks different from the training set (*e.g.* sandy road, snow, or painted ground). Motion predictive model restricts the application to robots that satisfy the predefined motion constraints (*e.g.* applying non-holonomic constraint on holonomic system will be detrimental).

Recently, dense optical flow received a lot of attention, with current state-of-the-art methods capable of producing on average, more than 85% of dense correspondences

less than 3 pixels error. Although the computed correspondences from optical flow are not very accurate, the dense nature of the correspondences helps mitigate the inaccuracy of individual correspondence. Another benefit of using dense optical flow is that it can avoid the shortcomings of typical sparse matches – in some scenes, the matches are clustered together around a small area of the image or encounter problem with planar degeneracy [Hartley and Zisserman, 2003], which may result in a biased motion estimate.

In this paper, we propose a novel dense optical flow method utilising the uncertainty measures which can enhance the visual odometry performance as well as the pose-graph SLAM. The contributions of this paper are threefold:

- Robust dense optical flow with estimated uncertainty
- A new weighted 8-points algorithm using uncertainty for the motion estimation
- Performance improvement of the monocular pose-graph SLAM

To the best of our knowledge, this is the first research work that utilises dense optical flow result for large motion visual odometry. The accuracy of the proposed visual odometry method is evaluated using the popular KITTI benchmark. The performance of our method is on par with other state-of-the-art methods while not relying on extra priors like vehicle's motion model, ground segmentation or keeping track of 3D scene points in multiple frames, making it suitable for general, long-term visual navigation.

2 Related Works

Monocular visual odometry can be categorised into two different approaches, namely direct methods and indirect methods. Direct methods (*e.g.* [Engel *et al.*, 2017]) optimise photometric error, which is also a cost commonly optimised in optical flow algorithms. Indirect methods optimise geometric error where sparse keypoint matches are usually used. Our proposed method is an indirect method where dense correspondences are treated like sparse keypoint matches. However, it also shares some similarity with direct method such that the dense optical flow method optimises for the photometric error.

The classical approach to estimate dense optical flow is obtained by optimising an objective (also known as energy or cost function) based on visual similarity and spatial smoothness (see (1)). This objective is traditionally optimised by iterative methods that update a single pixel's optical flow from an initial estimate [Horn and Schunck, 1980; Sun *et al.*, 2014], making those methods prone to solution getting stuck at local minima.

The earliest approach in using the full space of the discrete cost volume for global optical flow estimation was presented by [Chen and Koltun, 2016]. The highly regular structure of the cost volume means most of the operations is parallelisable for faster computation. The operation on the full space of the cost volume also avoids the solution from getting stuck at local minima. More recently, [Xu *et al.*, 2017] uses learned feature descriptor function, efficient semi-global matching technique, and homography fitting to obtain state-of-the-art result in both Sintel and KITTI dataset.

Past optical flow work also explored the addition of epipolar constraint to improve the optical flow result, both as a soft prior [Valgaerts *et al.*, 2008; Wedel *et al.*, 2009] or hard prior [Kitt and Lategahn, 2012; Yamaguchi *et al.*, 2013]. The latter approach restricts the two-dimensional search to one dimension. This greatly simplifies the computation, but dynamic objects do not satisfy this epipolar constraint, potentially making the result worse for dynamic part of the image. Recently, [Wulff *et al.*, 2017] proposed a new method that segments rigid scene from dynamic objects and applies rigidity constraint to improve optical flow of static regions of the image. This however, requires additional training step to obtain accurate static-dynamic segmentation. Similar to past work, we also apply epipolar constraint as a soft prior to improve the result for low texture, static region of the scene.

In contrast to sparse feature matching, dense optical flow methods typically do not provide sub-pixel accurate correspondences. However, the dense nature of the optical flow allows the error to be averaged out. A closely related literature to our work is [Valgaerts *et al.*, 2012] where a comparative study of sparse and dense correspondences is used to estimate the Fundamental matrix. Their work does not include the estimation of the uncertainty of the optical flow, and their experiments are focused on rigid scenes with small camera displacement. On the other hand, our proposed method is suitable for large translational motion, where there are dynamic objects in the scene, and large regions of the image have severe occlusion and non-overlapping regions between two image frames.

In general, monocular visual odometry can only estimate the camera motion up to an unknown translational scale factor. The translational scale can be recovered either by known dimensions about the environment (*e.g.* Pipe diameter for pipe crawling robot [Hansen *et al.*, 2011]), known motion constraints of the robot (*e.g.* non-holonomic constraint [Fanani *et al.*, 2017; Scaramuzza *et al.*, 2009], or by using fixed camera height assumption [Fanani *et al.*, 2017; Song *et al.*, 2016]). We argue that using known dimensions of the environment or known motion constraint on the robot/vehicle makes

the visual odometry less general, and thus only uses the fixed camera height assumption in our approach. Due to the robustness of our method, unlike [Fanani *et al.*, 2017; Song *et al.*, 2016], we only use a single cue (*i.e.* reconstructed ground plane) to estimate the camera height.

3 Dense Optical Flow

Classical optical flow algorithm optimises a cost function of the form

$$C(\boldsymbol{f}) = C_{data}(\boldsymbol{f}) + \lambda C_{reg}(\boldsymbol{f}) \tag{1}$$

where $\boldsymbol{f}$ is the computed optical flow, C_{data} is the data term that penalises visually dissimilar pixel, C_{reg} is the regularisation term that encourages spatially smooth variation of optical flow field, while λ controls the trade-off between the two terms.

We can represent the discrete matching cost of a set of candidate pixels in the second image to a corresponding pixel in the first image using a two-dimensional (2D) cost slice. Each pixel in the first image has their respective 2D cost slices, such that the full cost volume is four-dimensional (4D). The full discrete cost volume can be used directly to compute dense optical flow [Chen and Koltun, 2016; Xu *et al.*, 2017]. The four-dimensional cost volume has the size of $M \times N \times D \times D$, where $M \times N$ is the scaled down (1/3) dimension of the input images, while D is equals to $2 * d_{max} + 1$, where d_{max} is the maximum pixel displacement between the two images.

The dense optical flow method we use for visual odometry task is a modified version of the direct cost volume optical flow from [Xu *et al.*, 2017]. Currently, their method is one of the most accurate monocular dense optical flow methods for both KITTI and Sintel dataset, with a short computational time of under 9 seconds. They have made the code public, along with the fact that their method directly operates on the full cost volume allows us to make the necessary modifications to improve the performance, and also obtain an estimate of the optical flow uncertainty. The following section details our modifications to the method.

3.1 Epipolar Constraint on Direct Cost Volume Optical Flow

In a lot of the driving scene from KITTI dataset, large areas of the image are covered by low textured surfaces (*e.g.* road, the wall of buildings), which makes the task of finding correct correspondences difficult. It is especially crucial for the road surface correspondences to be correct as it is the primary means of fixing the translational scale in a monocular visual odometry.

One way to reduce the ambiguity of the matching is by applying epipolar constraint into the cost function in (1). We encourage the correspondences to be close to the epipolar line by increasing the cost of finding a match far from the line. This is accomplished as follows. First, Shi-Tomasi corner feature tracked by Kanade-Lucas-Tomasi (KLT) algorithm [Shi and Tomasi, 1994], are used as sparse correspondences for the 8-point algorithm to obtain an initial estimate of the Fundamental matrix. A truncated L2 cost is added to the cost volume to enforce the epipolar constraint. This is illustrated as follows.

Given fundamental matrix F, the epipolar cost C_{epi} of homogeneous coordinate of the pixel $\boldsymbol{x_1}$ in the first image being matched to pixel $\boldsymbol{x_2}$ in the second image is

$$C_{epi} = min\left(\alpha \frac{\langle \boldsymbol{x_2}, \boldsymbol{l} \rangle^2}{l_1 + l_2},\ \beta\right) \tag{2}$$

where $\boldsymbol{l}$ is the epipolar line equation, which is equals to $F\boldsymbol{x_1}$; $\langle \cdot, \cdot \rangle$ is the vector dot product; l_1 and l_2 are the first and second element of vector $\boldsymbol{l}$ respectively; α is the confidence factor (analogous to 1/variance of a Gaussian distribution); while β is the maximum (truncated) value of the epipolar cost.

Truncated L2 cost is used to enforce the epipolar constraint because a typical driving scene also contains independently moving objects (*e.g.* cars, pedestrians). When the pixel in the first image corresponds to a static point of the scene, the cost of finding the match far away from the epipolar line is increased proportionately to squared distance. Conversely, when a pixel in the first image corresponds to a point on a moving object, a truncated cost is applied, avoiding wrong matches to be found that satisfies the epipolar constraint but is visually dissimilar.

The epipolar constraint is added to the cost function before regularisation is applied. Figure 2 shows an example of the epipolar constraint being added to one of the cost volume slices.

3.2 Uncertainty Estimation

The direct cost volume processing of [Xu *et al.*, 2017] optical flow (DCFlow) allows the uncertainty of each correspondence to be easily estimated. One way to estimate the uncertainty is by assuming the cost is distributed following negative logarithm of a unimodal, bivariate Gaussian distribution. Then, the uncertainty is extracted by fitting negative logarithm of a Gaussian distribution to the cost. For a one-dimensional Gaussian distribution, this is just a quadratic curve fitting. An example is illustrated in Figure 3.

A simpler method to capture the uncertainty of negative logarithm of a Gaussian distribution (cost function) is by determining the width of the distribution at a fixed height. Note that the cost function must be normalised before the width is determined, such that the minimum cost has a value of zero. Then, the cost function can be expressed as a quadratic function of the form

Figure 2: Example illustrating epipolar constraint added to a cost slice (before spatial smoothness regularisation step) at varying strength. From left to right: first image with a pixel marked by a green star, second image with a bounding box enclosing the candidate matching pixels for the corresponding pixel marked in first image, cost slices representing the matching cost of corresponding candidate matching pixels with no epipolar constraint, truncated cost at 5, and truncated cost at 15 respectively. Note that the candidate matching pixels outside the boundary of the image is assigned a fixed cost (blue colour at the bottom of all cost slices).

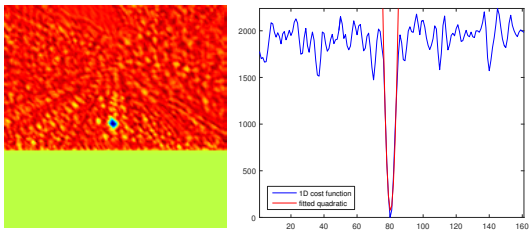

Figure 3: An example showing the fitting of negative logarithm to a cost slice (after spatial smoothness regularisation step). From left to right: 2D cost slice, a cross section (row) of the cost slice showing the original cost curve in blue and fitted quadratic equation in red.

$(x-\mu)^2/(2\sigma^2)$. At a fixed height h, the width of the cost function is then equals to $\sigma\sqrt{8h}$. Thus, the width is proportional to the uncertainty of the distribution.

A bivariate Gaussian distribution can also be understood as a combination of two one-dimensional Gaussian distribution perpendicular to each other. The general form of bivariate Gaussian's covariance matrix can be written as shown in (3) [Brooks *et al.*, 2001].

$$\Lambda = \kappa^2 R_\theta \begin{bmatrix} \gamma & 0 \\ 0 & 1-\gamma \end{bmatrix} {R_\theta}^T \tag{3}$$

where $\kappa^2 = {\sigma_x}^2 + {\sigma_y}^2$, $\gamma = {\sigma_x}^2/({\sigma_x}^2 + {\sigma_y}^2)$, and R_θ is the anti-clockwise rotation by angle θ.

We can simplify the uncertainty estimation by ignoring the directional information of the uncertainty while capturing only the square-root of the sum of squared of the uncertainty in both x and y directions (κ in (3)). This removes the computational requirement of Gaussian fitting and reduces memory requirement by a factor of four (covariance matrix of bivariate Gaussian has dimension 2×2). This essentially encodes the maximum error magnitude of each of the matching and is done as follows.

Given the cost of finding a matching pixel in the second image for a pixel $\boldsymbol{x_1}$ in the first image, the most likely matching pixel is obtained by finding the corresponding pixel with the lowest cost. Similar to the one-dimensional case, we need to normalise the cost volume by subtracting the minimum of individual cost slices (each cost slice corresponds to the two-dimensional matching cost of a pixel in the first image to a set of candidate pixels in the second image). Normalising the cost slices is equivalent to setting the probability of the matching pixel with the minimum cost to be equals to one.

Then, we can use the normalised cost to calculate the uncertainty magnitude, by determining the dimension of the largest rectangular bounding box that encloses all costs with the value less than a threshold v_c. The threshold (height) is tuned such that the bounding box only encloses a small region around the minimum and ignore random noise in the cost (as can be seen from Figure 3). If the elements of a vector $\boldsymbol{b}$ is made up of the width and height of the rectangular bounding box, the uncertainty magnitude is then equals to the Euclidean norm of the vector $\boldsymbol{b}$.

Similar to most top performance optical flow method, DCFlow has a post-processing step to remove unreliable matches from the semi-dense correspondences before EpicFlow [Revaud *et al.*, 2015] interpolation. This is accomplished by computing the forward and backward optical flow, and removing those matches that do not satisfy the forward-backward consistency. This post-processing step changes the uncertainty estimate such that the correspondences that got removed should be assigned a high uncertainty. We replace those values with the maximum uncertainty of the current optical flow estimate.

These provide us with a single channel, floating-point image encoding the uncertainty magnitude for every pixel correspondences for the scaled-down pair of RGB input images. We can scale the uncertainty image back to the original resolution by applying an image resize operation, followed by a multiplication of three (image rescaling factor). The estimated uncertainty can also be used to determine if the two input images are visually similar, which will be helpful when computing the loop

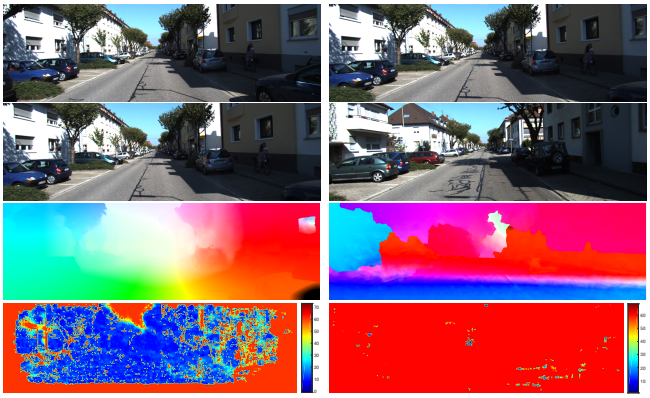

Figure 4: Example of estimated optical flow and uncertainty magnitude. From top to bottom: first input image, second input image, optical flow, uncertainty magnitude. Left column corresponds to sequential images, while right column corresponds to two input images with high structural similarity (SSIM) index, but is not of the same scene.

closure constraints (section 5). This is illustrated in Figure 4.

4 Inter-frame Motion Estimation

The dense optical flow and its uncertainty are used to estimate the inter-frame motion. The dense optical flow correspondences are treated like conventional sparse features matches, while the uncertainty is used during the sampling step of RANSAC, and also to apply a weighting to their corresponding equation in our weighted normalised 8-points algorithm. Ways to determine and handle small motion in the video sequences are also discussed in the following subsections.

4.1 Weighted 8-point Algorithm

We propose a new algorithm to estimate inter-frame motion based on the well-known normalised 8-points algorithm [Hartley, 1997]. Dense optical flow correspondences are treated similar to sparse feature matches in conventional 8-points algorithm. In the RANSAC step, the uncertainty of the optical flow is used to guide the sampling of the matches by increasing the likelihood of selecting correspondences (or matches) with a lower uncertainty. This decreases the required number of iterations to ensure a good inlier set is selected.

The second, and arguably the most important use of the optical flow uncertainty is to apply a weighting to each equation during the refinement step of the weighted 8-point algorithm. This ensures that the solution of the Fundamental matrix tends to fit better to those matches with lower uncertainty. The weight for i^{th} correspondence, w_i can be calculated as follows.

$$w_i = 1/(\kappa_i + \epsilon) \tag{4}$$

where κ_i is the computed uncertainty magnitude for the i^{th} correspondence, and ϵ is a small number (set to 10^{-10}) to prevent division by zero.

In the refinement step of the Fundamental matrix estimation, we also apply an additional reweighing step such that the Sampson distance is minimised. The Sampson distance is the first order approximation of the geometric distance between a pixel and its corresponding epipolar line for a given Fundamental matrix estimate [Hartley and Zisserman, 2003]. The reweighing is performed for three iterations to ensure convergence. Given an initial Fundamental matrix estimate F, homogeneous coordinates of matching pixels in both images $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$, the weighing function for the Sampson distance, ϕ_i can be calculated as follows.

$$\phi_i = \frac{1}{(F\boldsymbol{x_i})_1{}^2 + (F\boldsymbol{x_i})_2{}^2 + (F^T\boldsymbol{x_j})_1{}^2 + (F^T\boldsymbol{x_j})_2{}^2} \tag{5}$$

where the notation $(\boldsymbol{v})_k$ is the k^{th} element of the vector $\boldsymbol{v}$.

The new weighted normalised 8-point algorithm is detailed in Algorithm 1.

4.2 Scale Estimation

Unlike stereo visual odometry, the translational scale cannot easily be recovered due to the lack of a reliable reference (*i.e.* fixed, known stereo baseline). Thus, we can only estimate the scale by relying on other assumptions. The most common assumption used by monocular visual odometry is that the camera moves at a fixed height from a roughly planar ground. This is a good assumption for video sequence captured from any ground-based vehicle or robot. The rare cases the assumption is not valid are when the observable ground is not completely planar (*e.g.* sidewalk, speed bump).

The scale is determined by fitting a plane through the 3D reconstructed points that are roughly parallel to the zx (forward-right) plane of the camera axis. Assuming the road is visible roughly in the middle of the image, we use the reconstructed points below the camera (y coordinate of the 3D points is positive), and not too far to the side (image coordinate x within half the image width from the centre) of the camera. Plane fitting provides us with a plane equation satisfying $ax+by+cz+d = 0$. The height of the plane with respect to 1 unit of inter-frame translation is then equals to $-d/b$. If the height of the camera, h is known (calibrated from training data), the scale of the inter-frame translation s can be computed as follows.

$$s = -\frac{b}{d}h \tag{6}$$

Algorithm 1: Weighted normalised 8-point algorithm for inter-frame motion estimation

Data: Dense correspondences $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ (from optical flow), uncertainty of correspondences $\boldsymbol{u}$, intrinsic camera parameters $\boldsymbol{K}$
Result: Inter-frame motion, $\boldsymbol{R}$ and $\boldsymbol{t}$ (translation has unknown scale)

Compute normalised correspondences $\hat{\boldsymbol{m}}_{\boldsymbol{1}}$ and $\hat{\boldsymbol{m}}_{\boldsymbol{2}}$;
for *r = 1:N* **do**
  Find random 8 correspondences (with higher chance of selecting correspondences with lower uncertainty);
  Estimate Fundamental matrix $\boldsymbol{F}$ (typical 8-point algorithm);
  Find number of inliers n (based on distance from epipolar line);
  if $n < n_{previous}$ **then**
    Set $n_{previous} = n$;
    Set $\boldsymbol{F_{est}} = \boldsymbol{F}$;
  end
  if $n > (log(1 - 0.9999)/log(1 - (n/n_{total})^8))$ **and** $r > 10$ **then**
    break;
  end
end
for *r = 1:3* **do**
  Refine $\boldsymbol{F_{est}}$ by weighing all inliers equations with their corresponding weighing factor ($w_i\phi_i$ from (4) and (5);
end
Denormalise $\boldsymbol{F_{est}}$ and fit to the closest rank 2 matrix;
Estimate Essential matrix $\boldsymbol{E}$;
Make $\boldsymbol{E}$ has singular values of [1,1,0];
Extract two possible rotations and translations from $\boldsymbol{E}$;
Use chirality constraint to find the correct $\boldsymbol{R}$ and $\boldsymbol{t}$;

4.3 Small Motion Handling

The use of scaled-down images (one-third the original scale) for dense optical flow estimation cannot guarantee the accuracy of the matches when the pixel translation between two images is too small. This occurs when the vehicle moves very slowly or stops completely, causing the motion estimation to be error-prone. Small translational motion estimation is a common problem in most monocular visual odometry. This is because small parallax between two images leads to difficulty in estimating both motion and structure accurately.

We determine if the inter-frame motion is big enough using two separate conditions. First, the Shi-Tomasi corner matches have a median displacement magnitude of at least 2.5 pixels. Secondly, the third quantile (75%) of the computed optical flow has a magnitude greater than 5. We can reduce the computational time by skipping images that fails the Shi-Tomasi corner displacement condition, before the dense optical flow is computed. If either of the two conditions is not met, the inter-frame motion is assumed to be zero (no rotation or translation), and the first frame is stored to be used to compute the subsequent motion. For example, when the motion between frame at t and $t+1$ is not big enough, frame t is stored and the motion is estimated using frame t and $t+2$. If the motion is again not big enough, the motion is estimated for frame t and $t+3$, and so on.

5 Loop Closure

Loop closure is possible when a previously visited location is revisited. We determine the candidate frames for loop closure in three steps. The first step is by selecting frames with their estimated poses to be less than a fixed (metric) distance away, while having a difference in frame index no less than 200. The minimum frame index difference is enforced to prevent finding too many candidates within neighbouring frames. We can further reduce the possible candidates by only finding candidate loop closure images for every 10 frames.

The second step is to determine which of the candidate loop closure images are valid, by using the structural similarity index (SSIM) [Wang *et al.*, 2004]. We discard any images that have a SSIM index less than a set threshold (experimentally set to 0.38), and keep a maximum of three candidate images with the highest SSIM index. Lastly, the dense optical flow between the images and their possible neighbours are computed. The estimated uncertainty is used to determine if the optical flow is reliable, and only compute their inter-frame motion when the percentage of matches with an uncertainty less than a set threshold is greater than 20% (see Figure 4). During the loop closure, the inter-frame motion estimation step also checks for the small motion conditions as discussed in previous sections. This provides us with close loop constraints that links temporally far away poses that are spatially close to each other.

Loop closure is accomplished by using the robust lin-

ear pose graph-based optimisation method from [Cheng *et al.*, 2015]. Similar to other pose graph SLAM, their method treats all poses of the vehicle or robot as vertices, and inter-pose constraints (*e.g.* odometry and close loop constraints) as edges. Their method is used due to the method's robustness of handling outliers in the loop closure constraints, and effectively discarding wrong edges, preventing incorrect convergence result.

6 Experiment

The dataset we used to verify the performance of our proposed algorithm is taken from KITTI benchmark. For optical flow evaluation, we use the flow 2015 dataset [Menze and Geiger, 2015], while for the odometry, we use the odometry dataset [Geiger *et al.*, 2012]. For both experiments, we use the monocular RGB images (*image_2* folder). In the odometry experiment, we assume the camera is $1.68m$ above the ground, with zero pitch.

Due to the post-processing part of the DCFlow code not made available, we can only verify the optical flow result before homography fitting is applied to the EpicFlow [Revaud *et al.*, 2015] interpolated results. Based on KITTI 2015 optical flow dataset, by applying our epipolar constraint on the cost volume, we achieved a 0.6% improvement in accuracy (in terms of less than 3 pixels endpoint error criterion). The improvement is small due to the epipolar cost parameters (*i.e.* α, β in (2)) being set very low to accommodate for dynamic pixels in the scene. However, we can visually observe a noticeable improvement in the optical flow estimation for the ground pixels, not reflected by the large (3 pixels error) KITTI accuracy metric. We also implemented a homography fitting step based on the description of their paper.

The uncertainty estimate for the dense optical flow is visually inspected, where it was observed that occluded, out-of-bound or textureless regions of the image have high uncertainty value (see Figure 4).

For visual odometry result, we compare our performance with existing methods, specifically VISO2-M [Geiger *et al.*, 2011], MLM-SFM [Song *et al.*, 2016] and PMO [Fanani *et al.*, 2017]. We selected a few of the available sequences that contain slow moving vehicle in an urban environment (sequence 00), fast moving vehicle on a highway (sequence 01) and vehicle travelling in a loop (sequence 06) to gauge the performance of our proposed method. The result is summarised in Table 1.

Note that VISO2-M [Geiger *et al.*, 2011] and MLM-SFM [Song *et al.*, 2016] methods fail to estimate the visual odometry for sequence 01 due to the highly repeated structures of the scene, which cannot be reliably matched by the sparse feature matching technique their methods employ. Figure 5 shows our estimated trajectory for the vehicle's motion.

From the estimated motion trajectory (Figure 5) and computed error from the ground truth (Table 1), we can observe that our proposed method achieved very accurate estimation of translation. This is achieved without using bundle adjustment, motion model, or ground segmentation used by other state-of-the-art methods.

The rotation estimate is not as accurate as PMO [Fanani *et al.*, 2017], due to the fact that our method does not utilise the highly restrictive motion model of the vehicle. Their method also has an extra refinement step (JET) where both the features location and motion are jointly optimised, which reduces the rotation error by more than 85%.

7 Conclusion

In this paper, we have presented a novel, pose-graph SLAM method that utilises our weighted eight-point algorithm using dense optical flow with uncertainty. The optical flow uncertainty is estimated in a principled manner where negative logarithm of a bivariate Gaussian distribution is fitted to the computed matching cost. Using RANSAC, a set of inlier correspondences are identified. The novel weighted eight-point algorithm then optimises the weighted Sampson distance of each pixel correspondence to obtain a robust inter-frame motion estimate. The translational scale is then computed using the reconstructed ground plane and the known camera height. These provide us with the visual odometry estimate (front-end). Based on the estimated visual odometry, loop closure images and their corresponding poses are computed, which are then used in the linear pose-graph optimisation method (back-end) to obtain the final pose estimation result.

Further improvements are possible for this work. These include the use of joint feature location and motion optimisation similar to [Fanani *et al.*, 2017], which will further reduce the rotational error. Another extension of this work is to use the full bivariate covariance of the optical flow correspondences. By incorporating the directional uncertainty of each correspondence, the visual odometry performance can be further improved.

References

[Bradler *et al.*, 2015] Henry Bradler, Birthe Anne Wiegand, and Rudolf Mester. The statistics of driving sequences – and what we can learn from them. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

[Brooks *et al.*, 2001] M. J. Brooks, W. Chojnacki, D. Gawley, and A. van den Hengel. What value covariance information in estimating vision parameters? In

seq	VISO2-M		MLM-SFM		PMO		Ours		Ours + LC	
	rot (deg/m)	trans (%)	rot (deg/m)	trans (%)	rot (deg/m)	trans (%)	rot (deg/m)	trans (%)	rot (deg/m)	trans (%)
00	0.0209	11.91	0.0048	2.04	0.0042	1.09	0.0118	2.39	0.0117	2.03
01	n/a	n/a	n/a	n/a	0.0038	1.32	0.0107	1.149	n/a	n/a
06	0.0157	4.74	0.0081	2.09	0.0044	1.31	0.0066	1.58	0.0054	1.05

Table 1: Comparison of visual odometry accuracy for VISO2-M [Geiger *et al.*, 2011], MLM-SFM [Song *et al.*, 2016], PMO [Fanani *et al.*, 2017] and our proposed methods (LC stands for loop closure) for selected KITTI dataset. Note that sequence 01 does not revisit a previously visited location, therefore, loop closure does not apply.

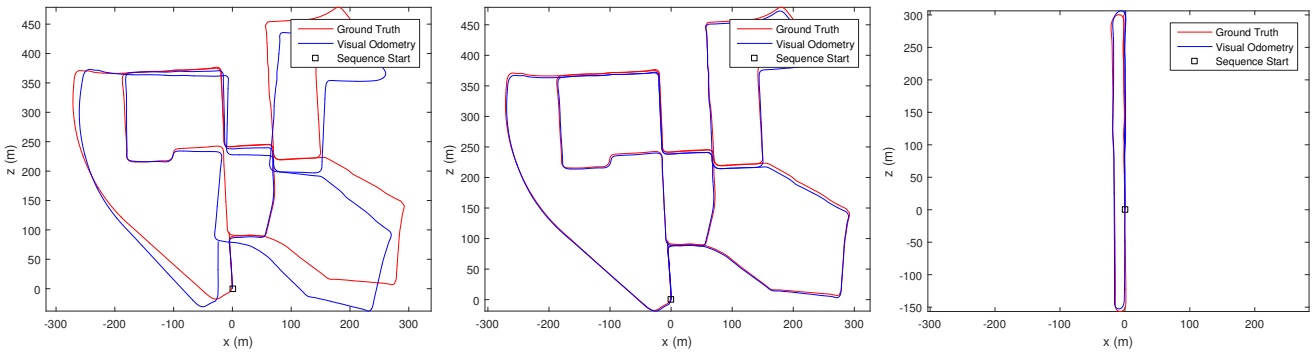

Figure 5: Comparison of the estimated motion trajectory and the ground truth motion. From left to right: Sequence 00 (without loop closure), 00 (with loop closure), 06 (with loop closure).

Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, volume 1, pages 302–308 vol.1, 2001.

[Chen and Koltun, 2016] Qifeng Chen and Vladlen Koltun. Full flow: Optical flow estimation by global optimization over regular grids. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4706–4714, 2016.

[Cheng *et al.*, 2015] Jiantong Cheng, Jonghyuk Kim, Jinliang Shao, and Weihua Zhang. Robust linear pose graph-based slam. *Robotics and Autonomous Systems*, 72:71–82, 2015.

[Endres *et al.*, 2012] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard. An evaluation of the rgb-d slam system. In *2012 IEEE International Conference on Robotics and Automation*, pages 1691–1696, May 2012.

[Engel *et al.*, 2015] J. Engel, J. Stckler, and D. Cremers. Large-scale direct slam with stereo cameras. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1935–1942, Sept 2015.

[Engel *et al.*, 2017] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[Fanani *et al.*, 2017] Nolang Fanani, Alina Stürck, Matthias Ochs, Henry Bradler, and Rudolf Mester. Predictive monocular odometry (pmo): What is possible without ransac and multiframe bundle adjustment? *Image and Vision Computing*, 2017.

[Geiger *et al.*, 2011] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *Intelligent Vehicles Symposium (IV)*, 2011.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[Handa *et al.*, 2014] A. Handa, T. Whelan, J. McDonald, and A. J. Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1524–1531, May 2014.

[Hansen *et al.*, 2011] P. Hansen, H. Alismail, P. Rander, and B. Browning. Monocular visual odometry for robot localization in lng pipes. In *2011 IEEE International Conference on Robotics and Automation*, pages 3111–3116, May 2011.

[Hartley and Zisserman, 2003] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer*

vision. Cambridge university press, 2003.

[Hartley, 1997] R. I. Hartley. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):580–593, Jun 1997.

[Horn and Schunck, 1980] Berthold K.P. Horn and Brian G. Schunck. Determining optical flow. Technical report, Cambridge, MA, USA, 1980.

[Kitt and Lategahn, 2012] Bernd Kitt and Henning Lategahn. Trinocular optical flow estimation for intelligent vehicle applications. In *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*, pages 300–306. IEEE, 2012.

[Kümmerle *et al.*, 2011] Rainer Kümmerle, Giorgio Grisetti, Hauke Strasdat, Kurt Konolige, and Wolfram Burgard. g 2 o: A general framework for graph optimization. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 3607–3613. IEEE, 2011.

[Meilland *et al.*, 2011] Maxime Meilland, Andrew Ian Comport, and Patrick Rives. Dense visual mapping of large scale environments for real-time localisation. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 4242–4248. IEEE, 2011.

[Menze and Geiger, 2015] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[Mur-Artal and Tardós, 2016] Raúl Mur-Artal and Juan D. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *arXiv preprint arXiv:1610.06475*, 2016.

[Revaud *et al.*, 2015] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1164–1172, 2015.

[Scaramuzza *et al.*, 2009] D. Scaramuzza, F. Fraundorfer, M. Pollefeys, and R. Siegwart. Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1413–1419, Sept 2009.

[Shi and Tomasi, 1994] Jianbo Shi and Carlo Tomasi. Good features to track. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 593–600. IEEE, 1994.

[Song *et al.*, 2016] S. Song, M. Chandraker, and C. C. Guest. High accuracy monocular sfm and scale correction for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):730–743, April 2016.

[Stachniss and Kretzschmar, 2017] Cyrill Stachniss and Henrik Kretzschmar. Pose graph compression for laser-based slam. In *Robotics Research*, pages 271–287. Springer, 2017.

[Sun *et al.*, 2014] Deqing Sun, Stefan Roth, and Michael J Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision*, 106(2):115–137, 2014.

[Tardif *et al.*, 2008] J. P. Tardif, Y. Pavlidis, and K. Daniilidis. Monocular visual odometry in urban environments using an omnidirectional camera. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2531–2538, Sept 2008.

[Valgaerts *et al.*, 2008] Levi Valgaerts, Andrés Bruhn, and Joachim Weickert. A variational model for the joint recovery of the fundamental matrix and the optical flow. *Pattern Recognition*, pages 314–324, 2008.

[Valgaerts *et al.*, 2012] Levi Valgaerts, Andrés Bruhn, Markus Mainberger, and Joachim Weickert. Dense versus sparse approaches for estimating the fundamental matrix. *International Journal of Computer Vision*, 96(2):212–234, 2012.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[Wedel *et al.*, 2009] Andreas Wedel, Daniel Cremers, Thomas Pock, and Horst Bischof. Structure-and motion-adaptive regularization for high accuracy optic flow. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1663–1668. IEEE, 2009.

[Wulff *et al.*, 2017] Jonas Wulff, Laura Sevilla-Lara, and Black Michael J. Optical flow in mostly rigid scene. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[Xu *et al.*, 2017] Jia Xu, Ren Ranftl, and Vladlen Koltun. Accurate optical flow via direct cost volume processing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[Yamaguchi *et al.*, 2013] K. Yamaguchi, D. McAllester, and R. Urtasun. Robust monocular epipolar flow estimation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1862–1869, June 2013.