Laser-Supported Monocular Visual Tracking for Natural Environments

Georges Chahine, Cedric Pradalier

Laser-Supported Monocular Visual Tracking for Natural Environments

Georges Chahine[1] and Cédric Pradalier[2]

***Abstract*— This paper presents and demonstrates a 2D laser-supported visual tracking solution, that can achieve reliable performance in unstructured scenes such as those seen in natural environment surveys. The method is shown to sufficiently stabilize scale and account for scale drift, as well as improve overall reliability. The suggested method is minimally invasive, does not require any additional parameter and does not necessarily require a laser, which can be replaced by any set of points with known depth, with no constraint on the temporal continuity of known points. We also test our method on 4 surveys, captured in a natural riverine environment, that proved to be challenging even for the state-of-the-art in visual tracking.**

I. Introduction

The practicality of using cameras for mapping and tracking is still driving research in related fields such as Visual Simultaneous Localization and Mapping (VSLAM), structure from motion and visual odometery. However, few methods focus on natural environments as demonstrated in [1].

Natural environments are challenging due to the lack of scene structure such as edges, features such as SURF[2] and ORB[3] and other geometrical shapes that might help stabilize visual tracking. Natural environments also suffer from fluctuating brightness conditions and in particular lens flare, as well as moving features such as tree branches and long grass. Also in [1], it was shown that tracking software based on the direct method *i.e*., that optimizes a photometric equation, tend to perform better than feature based methods while working with unstructured scenes. However, even the state-of-the-art in direct methods such as Direct Sparse Odometry (DSO) [4] poorly performs in unstructured scenes [1].

To the best of the authors' knowledge, there is no specialized VSLAM method for natural environment, and the existing solutions are not always reliable [1]. This paper does not offer a new VSLAM approach, rather, proposes a minimal support system for visual tracking that enables it to be sufficiently robust for subsequent use.

Most existing laser-supported VSLAM solutions such as [5], [6] are based on the availability of 3D lasers, which are costly and sometimes bulky to carry in natural environments. Sheng *et al*. [7] also comes short at solving the presented problem, since PTAM [8] is feature based and therefore unsuitable for natural environments [1].

[1]Georges Chahine is with the College of Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA gchahine@gatech.edu

[2]Cédric Pradalier is with the College of Computing, Georgia Institute of Technology, UMI 2958 GT-CNRS (Georgia Tech Lorraine), Metz, France cedric.pradalier@georgiatech-metz.fr

More closely related is the work of Zhang *et al*. [9], suggesting a 2D laser fusion approach to recover scale with a monocular camera. The main disadvantage of the proposed method is the requirement for reconstruction of a semi-dense surface around laser points, making the method unsuitable for unstructured scenes.

The proposed method is a minimally invasive modification to DSO, for the purpose of using 2D laser measurements. DSO was specifically chosen based on a previously conducted survey [1] that compared the state-of-the-art in VSLAM, and found that DSO is most suitable candidate for unstructured natural scenes. The proposed method does not introduce any new parameter to the existing VSLAM structure. Further detailed in III-A, the laser is assumed to be installed in the normal direction of the movement, meaning no assumption of overlapping laser points between camera poses. This work however assumes an overlap between the camera field of view (FOV) and the laser range, yet it is not critically dependent on laser measurements as they can be interrupted at any time, such as when DSO is unable to track projected laser points, with no immediate effect on stability.

In the remainder of this paper, we show the different steps undertaken to stabilize visual tracking with a focus on natural environments, with the purpose of recovering translational scale and accounting for scale drift, as well as improve overall reliability.

II. System

In this section, we detail the steps undertaken to include laser measurements with the purpose of stabilizing visual tracking. As shown in Fig. 1, the system is dependent on a camera feed and a sparse disparity image, here generated from a 2D laser. Still in the same figure, it is assumed that the extrinsics are inferred from a CAD model, further detailed in section III.

A. Laser Projection

We solve the laser projection problem using the classic pinhole camera model. Given a 2D laser ray that passes through the FOV of a camera, we project the laser point onto the corresponding camera image according to [10]:

$$P = A[R|t]L \tag{1}$$

Where A, $[R|t]$ and L respectively represent the camera intrinsics, the geometric transformation from the laser plane to the camera image plane (extrinsics), and the 3D coordinates of the laser points. Depth of the laser points in the image plane is calculated by taking the third component of

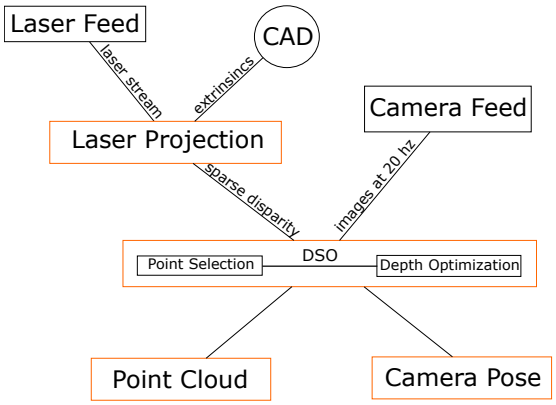


Fig. 1: System data flow

the product of the last two terms of the above equation, as in $[R|t]L$. The laser point coordinates u, v in the image frame are then recovered such as $[u, v, 1] = \frac{[p_x, p_y, p_z]}{p_z}$ with $P = [p_x, p_y, p_z]$.

Subsequently, we publish a sparse disparity image (Fig. 2) using the previously projected laser coordinates and their corresponding depth values in the camera frame.

B. Laser-Supported Visual Tracking

Visual tracking algorithms must have a point or feature selection scheme, the nature of the latter being largely dependent on the method being direct or indirect. We have already established in Section I that direct methods are more suitable for natural environments, we therefore propose the following:

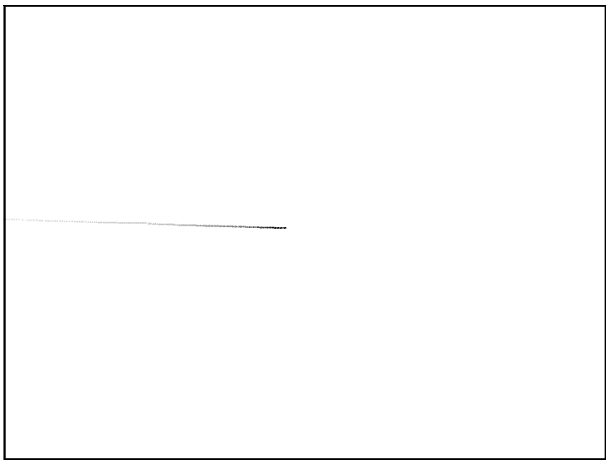
1) Forcing the selection of projected laser points as candidates to be tracked
2) Modifying the optimization step to enforce the known prior from the disparity image

Forcing the selection of laser points is recommended to improve the likelihood of tracking laser points, with no expectation that visual tracking will keep track of all the selected points. The selection process however varies from algorithm to another, and it might be harder to force point selection in feature based methods such as ORB SLAM[11], compared to direct methods that inherently require the selection of pixels for subsequent photometric optimizing.

If the received disparity image is empty, visual tracking continues to run but becomes vulnerable to accumulating scale drift and prone to failures. Once few laser points are tracked again, the scale is recovered and any built-up scale drift is eliminated. Other non-laser points are scaled according to the average scale inferred from the tracked laser points in other words, each time a new laser point is tracked, the inferred scale s for other non-laser points is updated such as:

$$s = \frac{1}{N} \sum_{n=1}^{N} \frac{d_t}{d_r} \tag{2}$$

Where N is the total number of tracked laser points, d_t is the real point depth inferred from projected laser measurements, and d_r is the VSLAM assigned depth, assumed to be random and captured before setting $d_r = d_t$ as previously discussed.



(a) Disparity image with inverted colors.

(b) Projected laser line on the actual scene.

Fig. 2: Laser line projected on the corresponding camera, using the pinhole model and the backpack frame CAD extrinsics 3.

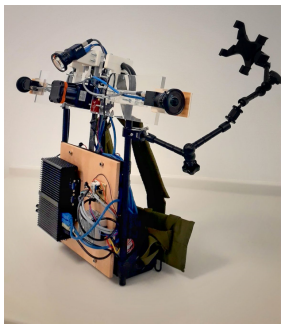
In addition to improving scale convergence time, the purpose of scaling non-laser points is to prevent depth discontinuities in-between laser and non-laser points, which might result in classifying either set of points as outliers.

III. Setup and Evaluation

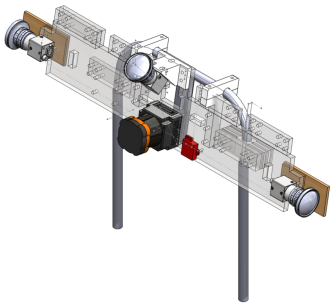
A. About the Dataset

The dataset is that of a riverine environment, with few snapshots shown in Fig. 4 and captured using the backpack shown in Fig. 3. The transformation matrix representing the extrinsics is calculated by using measurements inferred from the CAD model, while also taking into account the location of the image plane inside the camera hull, the latter provided by the camera manufacturer datasheet.

The testing data is mostly a combination of open fields, bushes, trees and long grass. In addition to the challenges



(a) Backpack used for data acquisition in the natural environment Image



(b) Backpack CAD model

Fig. 3: Survey recording was made using the backpack shown above. All fixtures were 3D printed or laser-cut according to a pre-designed CAD model, so that to minimize geometric projection error in-between components

associated with natural environments, the scene (camera angle) (Fig. 9(b)) is tilted in favor of the narrower vertical direction, meaning less room is available for feature tracking compared to a horizontally held camera.

The testing data for this paper consists of 4 surveys captured on a site near Nancy, France. In total, the surveys consist of over 1700 seconds of camera footage and laser recordings captured at 20 frames per second and 40 Hertz, respectively.

B. Laser-supported DSO

As previously discussed, the projected laser line is communicated as a disparity image shown in Fig. 2.

We therefore implement a laser projection node based on the method suggested in section II, that uses the Robot Operating System (ROS) to subscribe to laser sensor readings and publish a disparity image.

Subsequently, we modify DSO source files so that it subscribes to the disparity image generated by the laser projection node. Once DSO receives a non-empty disparity image, it will force the selection of the corresponding points in its point selection step. Given that the geometric transformation from laser to camera is subject to noise such as 3D printing precision uncertainties, we also select neighboring pixels to account for such errors and to improve the likelihood of tracking a laser point. Finally, during the point optimization step, tracked laser points are scaled using the laser disparity image, while remaining non-laser points are scaled as discussed in section II.

C. Discussion

1) Quantitative Assessment: Initially motivated by the need for a support system for DSO, laser-supported DSO proved to be a reliable solution, as shown in the completion rates in Fig. 5.

Completion rates were generated by dividing the tracked survey time before complete failure over the total survey time. In details, each survey was divided into 4 equal parts, for a total of 16 parts. Each part was subsequently evaluated against the two tracking methods, while allowing a maximum of 3 attempts for each part. Finally, The final tracked time for a given survey is the sum of all the tracked portions of each of its 4 parts. Given that most robotic systems involve more than one sensor, usually at least a GPS and an IMU, robustness in terms of completeness for visual surveys is highly desired, even if the results are noisy or biased. Many techniques, such as factor graphs, are available in literature and can greatly benefit from camera input to refine position, attitude and even scale estimation.

Scale drift is a major impediment often faced when working with monocular visual tracking solutions, and DSO is particularly prone to drift [1]. Fig. 6 shows the difference between native DSO behavior and laser-supported DSO in terms of scale drift. The figures show that laser-supported DSO is capable of recovering and maintaining correct scale throughout the survey. Note that a scale value in the vicinity of 1 means correct scale estimation. The ground truth data was generated by calculating the total trajectory length from an on-board inertial navigation system (INS), and assuming constant speed movement in a rolling $\sim$5 seconds window. For the same survey, Native DSO shows a highly nonlinear scale trend. Still in Fig. 6, ripples and peaks are localized violations of the constant speed assumption, used to generate the ground truth. An example of these violations is when the author was carrying the sensor backpack (Fig. 3) and had to navigate through muddy or rough terrains, sometimes pausing for few moments to perform a manoeuvre. The latter statement is reinforced by observing that both native and laser-supported DSO report localized scale jumps around the same frame number (400, 600, 700 and 1200), and by manual inspection of survey footage. The terrain corresponding to Fig. 7(a) and Fig. 7(b) was easier to navigate and therefore, ripples are less nuanced. Even-though both methods were able to fully track this particular survey, map and pose are evidently better suited for post processing when correctly scaled.

Laser-supported DSO is also capable of recovering correct scale after tracking is temporarily lost: Fig. 7(c) shows an incident of tracking loss around frame 2000, a common occurrence from which DSO typically recovers, yet at the expense of a considerable jump in scale. Still in the same figure, laser-supported DSO is shown to recover correct scale estimation as tracking resumes after non-catastrophic failure. Fig. 7(a) and Fig. 7(b) show correct scale estimation over 5000 and 2500 keyframes, respectively.

Fig. 8 shows performance metrics for angular drift using both native and laser-supported DSO. Fig. 8(b) shows that laser-supported DSO suffers 30% less drift as of 2000 keyframes (around 300 seconds). Fig. 8(a) shows that laser-based DSO takes more time to converge, yet locally exhibits a good follow-up on heading change increments. Still in the same figure, note the sharp turn of around 90 degrees undertaken around keyframe 1200, causing an angular drift in both laser-supported and native DSO, with the former

(a) (b) (c)

Fig. 4: Sample snapshots from natural environment surveys captured nearby a riverbed near Nancy, France. The same data is used for evaluation in section III.

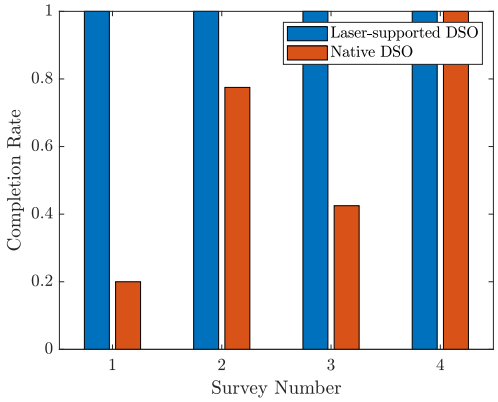


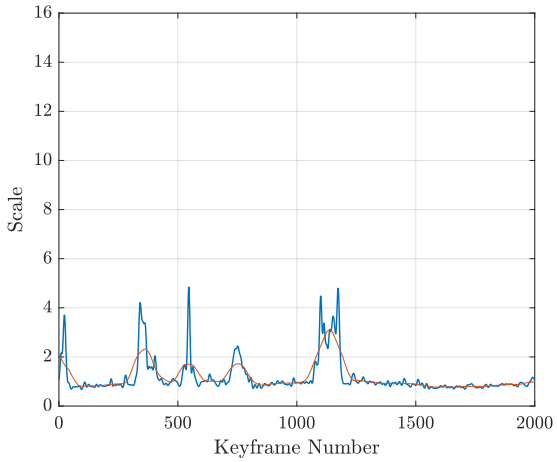
Fig. 5: Completion rates for native (unmodified) DSO and laser-supported DSO in 4 natural environment surveys.

exhibiting a modest recovery compared to the latter.

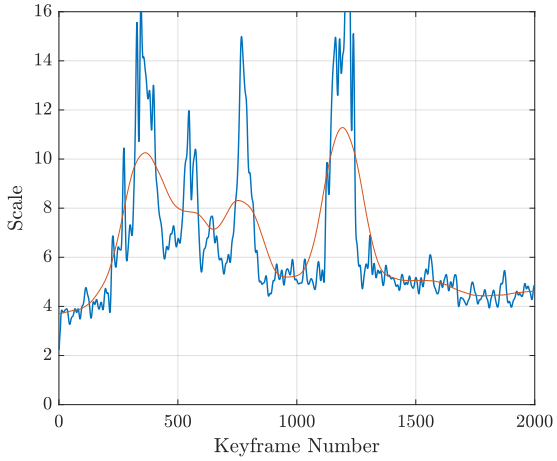
The results for heading estimation however, are not as impressive as the ones previously shown for translational scale. This is due to the fact that a projected 2D laser line does not, by itself, offer sufficient information to constrain a 6 DOF transformation. Nevertheless, DSO has been shown in Fig. 9 to be capable of tracking previously detected laser lines, therefore constraining rotations to a limited extent as long as several laser lines are tracked. Unfortunately, the traceability of previously projected lines is less likely to happen on rough turns, due to motion blur and the partial failure of the underlying tracking model.

It has been shown that the potential for recovery of translation scale can occur once few laser points are tracked again, the same cannot be said on angular/heading drift. This is due to the inherent temporal nature of rotations, namely those that can cause visual tracking to fail or drift *i.e.*, a rough turn event will degrade the quality of the tracking, yet once such rotation is complete, the information is permanently lost since no further data on the completed rotation will become available in the future.

2) Qualitative Assessment: The snapshots shown in Fig. 9 show a sample run of laser-supported DSO on one of the surveys, while Fig. 9(b) shows that DSO was able to track

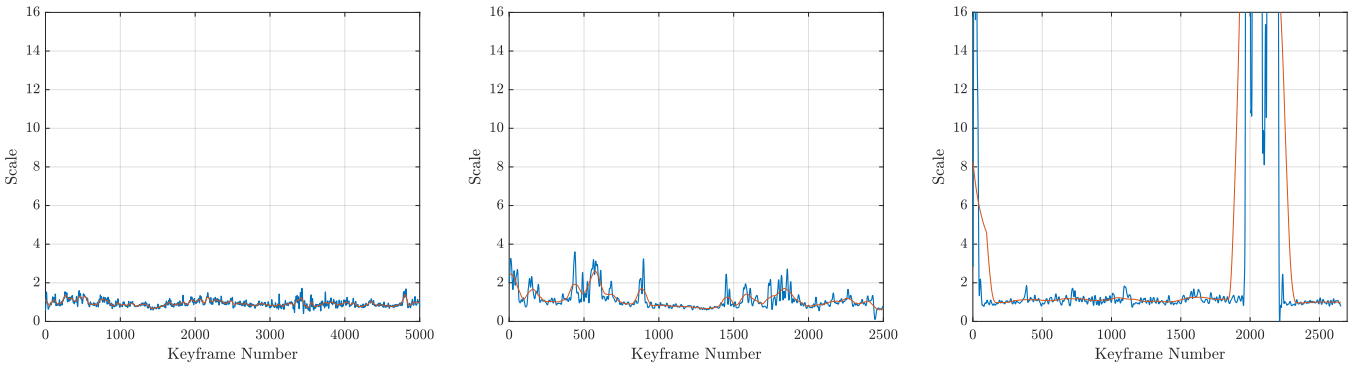


(a) Laser supported DSO scale metrics



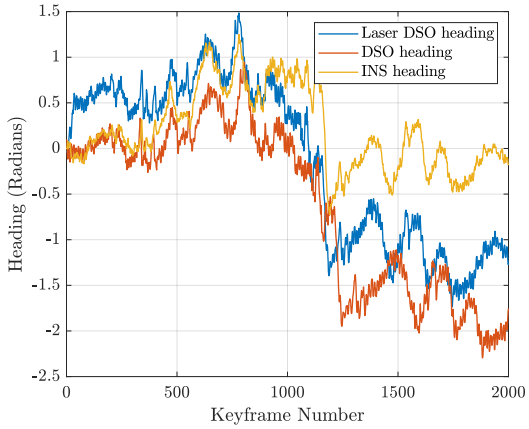
(b) Native DSO scale metrics

Fig. 6: Scale evaluation for both laser-supported DSO and native DSO for the only survey that both native and laser-supported DSO completed without failure.

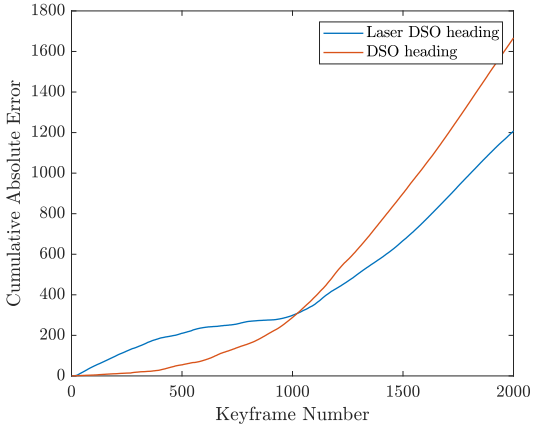


(a) Survey 1 (∼700 seconds) (b) Survey 2 (∼350 seconds) (c) Survey 3 (∼400 seconds)

Fig. 7: Scale estimation for 3 different natural environment surveys using laser-supported DSO, compared to ground truth. A scale value in the vicinity of one means correct scale estimation



(a) Heading comparison with INS ground truth



(b) Cumulative absolute error

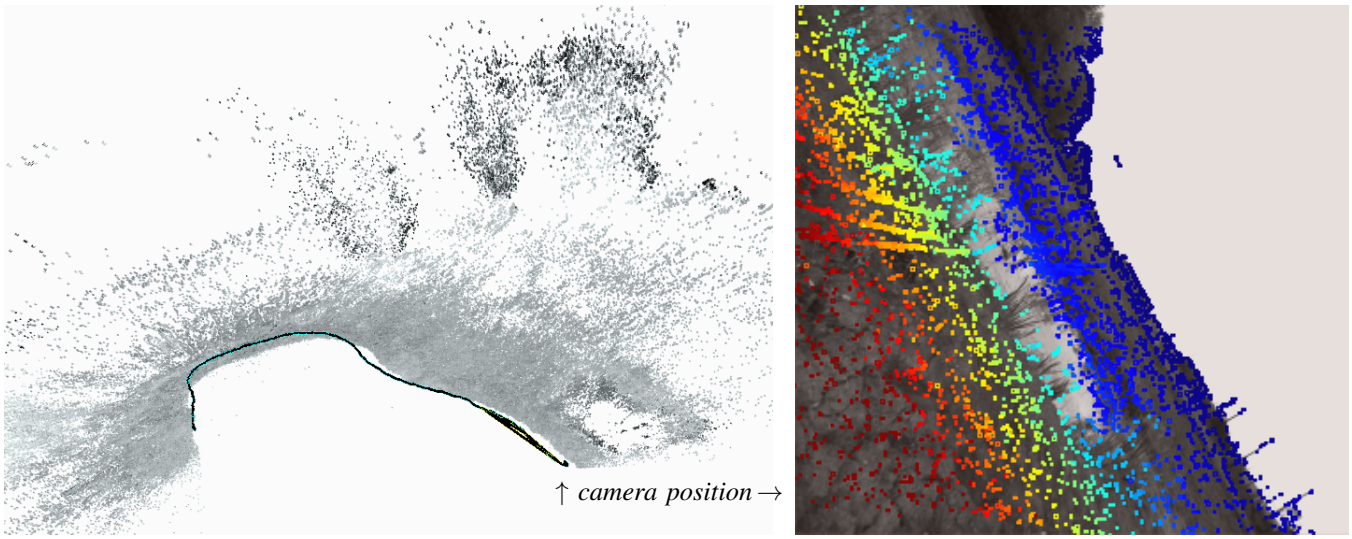
Fig. 8: Comparison of native and laser-based DSO in terms of angular drift. The ground truth is inferred from the on-board Inertial Navigation System (INS)

laser lines from previous frames.

The maps generated by DSO such as the one shown Fig. 9(a) show an improved visualization of the point cloud due to the inclusion of laser points and drift compensation. The projected laser lines such as the ones shown Fig. 9(b) also seem to propagate into the scene depth (out of the camera lens), providing a well distributed depth line that spans as far as 10 meters into the scene. Still in the same figure, notice the absence of tracked point inside the water pod, given that the laser beam is completely absorbed by the water, and the apparent blank hole left in the point cloud at the corresponding camera position.

Further, it was noticed while running the surveys that occasional loss of laser tracking can occur, due to lack of image information such as overexposed scenes and fast rotations. Passing nearby water pods causes the laser ray to be temporarily lost as it is dissipated in the water. Such effects contribute to the ripples seen in Fig. 7, but did not affect the overall stability of laser-supported DSO.

While no new parameter was introduced, laser-supported DSO appears to be sensitive to already existing parameters, such as the total number of tracked points. Typically, increasing the number of tracked points (above 4000 points) will improve robustness however, it was observed that doing so as laser-supported DSO is still initializing will occasionally prevent the proposed method from converging to the depth inferred from laser points, as it ignores the depth estimates as outliers and reverts to its original behavior. There was no issue in increasing the number of tracked points after DSO converges, typically few seconds after initialization. Increasing the maximum number of tracked keyframes in DSO favorably affected the performance of laser-supported dso, as previously projected laser lines become abundantly available and help constrain motion in subsequent frames.



(a) Laser-supported DSO point cloud (inverted colors), showing a muddy terrain with a water pod, and few far trees

(b) Sparse depth image showing few tracked laser rays

Fig. 9: Snapshots taken from laser-supported DSO, showing a point cloud and the corresponding image at the current position.

IV. CONCLUSIONS

We have presented a minimally invasive solution for the-state-of-the-art in visual tracking, enabling such methods to reliably perform even in the most challenging environments. This work is relevant to both roboticists and environmentalists, given the absence of specialized method in literature that can handle the complexity of unstructured natural scenes. As expected, knowing the depth of few features/pixels can help stabilize visual tracking and propagate the change to the remaining unknown points.

The generated map and pose from laser-supported DSO is far from perfect, yet opens the door for further refinement and draws some attention to the lack of visual tracking techniques in unstructured scenes. Future work will focus on fusing map and attitude from multiple cameras as well as the lidar, using similar datasets taken at difference times of the year. Finally, we aim at achieving temporal alignment of 3D maps generated in the natural environment, with the purpose of automatic quantification of long-term natural changes: a valuable asset for environmentalists seeking to easily yet reliably monitor slow natural changes.

ACKNOWLEDGMENT

This work was funded in part by the Grand Est Region, the Zone Ateliers Moselle, as well as the Rhine-Meuse Water Agency in France.

REFERENCES

[1] G. Chahine and C. Pradalier, "Survey of monocular slam algorithms in natural environments," in *2018 15th Conference on Computer and Robot Vision (CRV)*, May 2018, pp. 345–352.

[2] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 404–417.

[3] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 2564–2571.

[4] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Apr 2017.

[5] P. Newman, D. Cole, and K. Ho, "Outdoor slam using visual appearance and laser ranging," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, May 2006, pp. 1180–1187.

[6] D. M. Cole and P. M. Newman, "Using laser range data for 3d slam in outdoor environments," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, May 2006, pp. 1556–1563.

[7] J. Sheng, S. Tano, and S. Jia, "Mobile robot localization and map building based on laser ranging and ptam," in *2011 IEEE International Conference on Mechatronics and Automation*, Aug 2011, pp. 1015–1020.

[8] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, Nov 2007, pp. 225–234.

[9] Z. Zhang, R. Zhao, E. Liu, K. Yan, and Y. Ma, "Scale estimation and correction of the monocular simultaneous localization and mapping (slam) based on fusion of 1d laser range finder and vision data," *Sensors*, vol. 18, no. 6, 2018. [Online]. Available: http://www.mdpi.com/1424-8220/18/6/1948

[10] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. New York, NY, USA: Cambridge University Press, 2003.

[11] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017. [Online]. Available: https://doi.org/10.1109/TRO.2017.2705103