Localizing People in Crosswalks using Visual Odometry: Preliminary Results

Marc Lalonde[1], Pierre-Luc St-Charles[1], Délia Loupias[2], Claude Chapdelaine[1] and Samuel Foucher[1]
[1]*Vision and Imaging Team, Computer Research Institute of Montreal, 405 Ogilvy Ave. suite 101, Montreal, Quebec, Canada*
[2]*Département de Génie Électronique et Informatique, Institut National des Sciences Appliquées, 135 Avenue de Rangueil, 31400 Toulouse, France*

Keywords: Visual Odometry, SLAM, Direct Sparse Odometry.

Abstract: This paper describes a prototype for the localization of pedestrians carrying a video camera. The application envisioned here is to analyze the trajectories of blind people going across long crosswalks when following an accessible pedestrian signal (APS), in the context of signal optimization. Instead of relying on an observer for manually logging the subjects' position at regular time intervals with respect to the crosswalk, we propose to equip the subjects with a wearable camera: a visual odometry algorithm then recovers the trajectory and spatial analysis can then determine to which extent the subject remained within reasonable boundaries while performing the crossing. Preliminary tests in conditions similar to a street crossing show that our results qualitatively agree with the physical behavior of the subject.

1 INTRODUCTION

The accurate 2D localization of deformable objects such as pedestrians without a top-down view or a planar scene assumption is a challenging task. In an unconstrained setting where other objects might be simultaneously moving around a target of interest, and where static visual references are few, most classic vision-based solutions are prone to failure. In this work, we propose an early prototype to measure the trajectory of blind subjects crossing a street intersection with the aid of various accessible pedestrian signals (APS). Our goal is to determine which signals are more adequate in terms of pitch, melody, etc. in guiding a blind person to align themselves with the crosswalk, and to remain within its boundaries throughout the crossing. For this, we measure each subject's deviation with respect to the center of the crosswalk, varying the signal used in each experiment. Better signals should, on average, minimize such deviations. Previous data collection protocols usually required researchers to visually estimate the deviations as the person walks in front of them, which is obviously inaccurate and labor-intensive (Laroche et al., 2000). Since hundreds of crossings may also be required for the proper statistical analysis of deviations in our problem, this approach is unsuitable. On the other hand, hardware-based localization solutions are not always adequate due to the spatial accuracy required for proper analysis ($\approx$15 cm); for example, the accuracy of consumer GPS devices is a few meters in good conditions. Furthermore, the equipment used should not disturb the subjects' progress during crossing, and measurements should be done inside a fairly large volume since the intersection that is selected for the experiment is a six-lane boulevard with a median (total walking distance is 30m).

Recent advances in robot vision, most notably in Simultaneous Localization and Mapping (SLAM) techniques, may provide an elegant solution to our problem: if the subject is wearing a camera while performing the crossing, the tracking of the camera pose would allow the recovery of the 3D trajectory of the subject, hence its deviation with respect to some reference points. This paper is organized as follows: first, Section 2 describes a previous approach to the problem and also provides some background information about visual odometry; in Section 3, the proposed strategy is exposed; and finally, in Section 4 we report on some preliminary results gathered during a short experiment.

Lalonde, M., St-Charles, P-L., Loupias, D., Chapdelaine, C. and Foucher, S.
Localizing People in Crosswalks using Visual Odometry: Preliminary Results.
DOI: 10.5220/0006646904820487
In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*, pages 482-487
ISBN: 978-989-758-276-9

2 BACKGROUND WORK

A vision-based approach to this problem has been proposed in the past (Lalonde et al., 2015). It relied on the post-experiment analysis of video footage captured using a handheld camera to determine the subject's movement from an offset point of view. The subject's feet were localized with respect to known landmarks (markings painted on the ground) for spatial referencing. Such an approach is convenient in terms of data acquisition: an observer merely needs to walk behind the subject with a camcorder, video acquisition and management is easy, resolution is always good, etc. However, many challenges made the analysis phase difficult, most notably the large variations in illumination and ensuing cast shadows, as well as the lack of robustness of the feet tracking algorithm. In addition, the method was dependent on the presence of several lines painted on the pavement, and their location had to be precisely known *a priori*.

In this paper, we tackle the movement mapping problem using a vision-based simultaneous localization and mapping (SLAM) approach. The idea behind this kind of approach is to use visual data to concurrently build a model of the local environment (i.e. a "map") and estimate the state (or location) of the camera within it. In our case, the map of the environment is not our primary focus, as our specific application only relies on odometry, and loop closure is not needed (i.e. we analyze one-way street crossings). Nonetheless, environment maps can be used to correct scaling issues found in monocular camera setups (as discussed further in Section 3).

SLAM methods can be separated into direct and indirect approaches. Indirect SLAM methods such as ORB-SLAM (Mur-Artal et al., 2015) and PTAM (Klein and Murray, 2007) typically use keypoint detectors to extract unique landmarks from the observed images, and then estimate scene geometry and camera extrinsics using a probabilistic model. This classic approach is quite efficient in practice due to the sparse nature of visual keypoints, and it is quite robust to noise in geometric observations. However, these keypoint-based methods fail when the observed images are composed mostly of uniformly-textured regions. Direct SLAM methods such as DSO (Engel et al., 2017) and LSD-SLAM (Engel et al., 2014) rely on local image intensities instead of sparse keypoints to represent observations in their model. The advantage of this approach is that it can use and reconstruct any observed surface with an intensity gradient. This is a crucial requirement for our application, as most street crosswalk surfaces show repetitive landmarks and high-frequency or uniform textures, which would hinder the performance of an indirect SLAM method. Besides, note that self-localization using only a camera has been studied extensively before, but mostly for robots or vehicles in large scale contexts (Se et al., 2002; Pink et al., 2009; Brubaker et al., 2016). In our case, a person's gait directly affects the stability and height of the camera, which can in turn hinder the performance of traditional localization methods based on landmarks or holonomic constraints.

For a more complete look at various SLAM methodologies and algorithms, we refer the reader to the recent survey of (Cadena et al., 2016).

3 STRATEGY

In this work, we take advantage of the recent developments in robot vision and SLAM, and explore the use of visual odometry techniques to localize a person during a street crossing. So, instead of having someone hold the camera behind the subject and try to track both the subject and the environment (using e.g. added markers on the ground for proper localization), we equip the subject with a calibrated camera facing the street. Localizing the subject then amounts to tracking the camera pose throughout the crossing.

As noted before, SLAM using a single camera setup (i.e. a monocular setup) entails that the absolute scale of the environment is unknown — this is a problem for us, as deviations need to be recorded and registered in a fixed coordinate system. Some SLAM extensions rely on GPS, IMUs, or altimeters to correct this issue via sensor fusion using Extended Kalman Filters (Lynen et al., 2013). Others instead rely on assumptions about the camera height above the ground plane (Song et al., 2016), or about its movement in very constrained settings (Gutiérrez-Gómez et al., 2012; Scaramuzza et al., 2009). In our case, we obtain camera trajectories using the Direct Sparse Optimization (DSO) method (Engel et al., 2017), and then fix this scaling issue by solving a camera Perspective-n-Point (PnP) problem using calibration boards placed around the crosswalk. Since we know the exact dimensions and grid layouts of these boards, we can determine their orientation and distance to the camera in specific key frames of the analyzed video sequences using the OpenCV calibration toolbox. These distances can then be averaged and used to properly scale the "map" provided by the SLAM algorithm. Furthermore, by fixing a calibration board directly on the ground, a coordinate space reference can be created, meaning all experiments can be registered to the same coordinate system. Finally, note that we could also use the length of the crosswalk

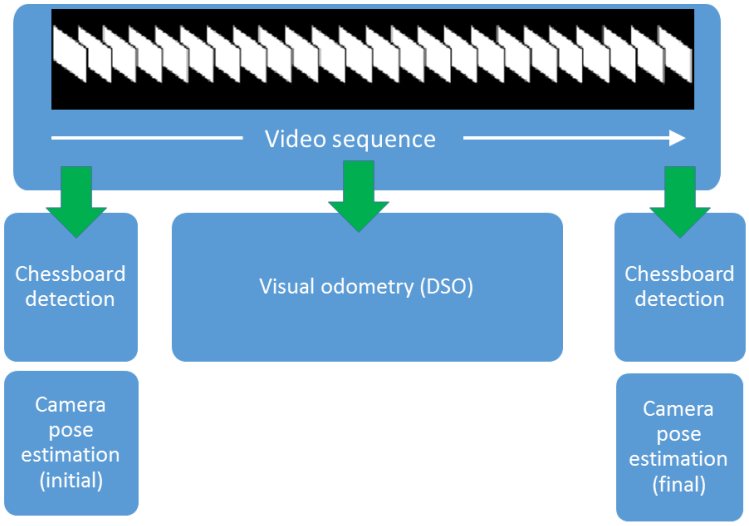

Figure 1: Block diagram of the approach.

(which is known *a priori*) to roughly validate the scale determined by solving the PnP problem. The strategy is depicted in Fig. 1.

4 RESULTS AND DISCUSSION

4.1 Experimental Results

Our preliminary experiments were conducted in a 15m x 5m zone of an outdoor parking lot, so as to simulate a street intersection (reduced by a scale of 1/2). The experimental setting is depicted in Fig. 2. We laid white tape on the pavement so as to form a 2m-wide corridor. A single subject was outfitted with a chest mount body harness equipped with a Hero3+ GoPro. The subject then simulated multiple crossings inside the corridor while carrying the GoPro, with varying trajectories with respect to the centerline of the corridor. The GoPro camera was oriented in portrait mode and slightly tilted toward the ground, so that both the horizon (including neighboring buildings, parked cars, street furniture, etc.) and the ground (pavement, line markings, etc.) were visible in the video frames. One key advantage with this camera configuration is the possibility of adding landmarks such as chessboard patterns on the ground, visible at the beginning as well as the end of the simulation. This allows the computation of the exact camera pose (3D position and orientation) at those moments, which corresponds to the initial/final absolute anchor points for the (relative) VO-computed trajectory, as mentioned in Section 3.

An example of a simulated crossing is given in Fig. 3, where we present some video frames as well as the top view of the 3D trajectory provided by DSO[1]. This top view representation is in line with the actual path followed by the subject: the starting point is in the middle of the corridor, there is a drift towards the right up to the edge of the corridor (roughly at mid point during displacement), and then a realignment towards the center.

4.2 Discussion

Overall, video sequences of nine crossings were collected. It however should be pointed out that only seven of them have been processed successfully by DSO: for the other sequences, the algorithm either lost track of the camera position mid-crossing, or it was unable to go beyond the initialization stage due to strong orientation variations in early frames. In that regard, an observation can be made about initialization. It seems that the first frames of the video sequence greatly influence DSO's behavior: if visual odometry starts as the person wearing the camera is already in motion, the initial estimate for the camera pose may be irreversibly biased, without any possibility of recovery. We hypothesize that the cause is due to the oscillatory nature of a person's walking pattern which induces an undesirable orientation of

[1]The implementation used in this work has been made available by (Engel et al., 2017) at https://github.com/JakobEngel/dso

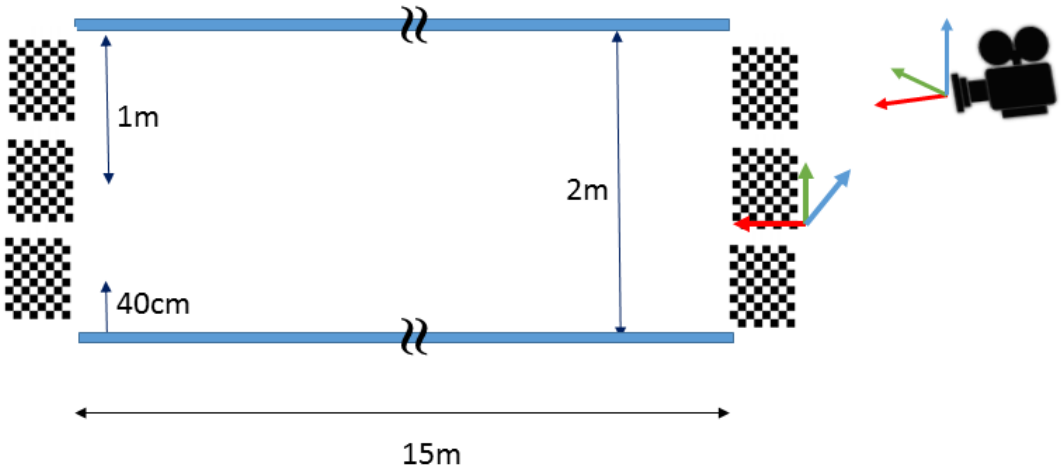

Figure 2: Sketch of the experimental crosswalk setup.

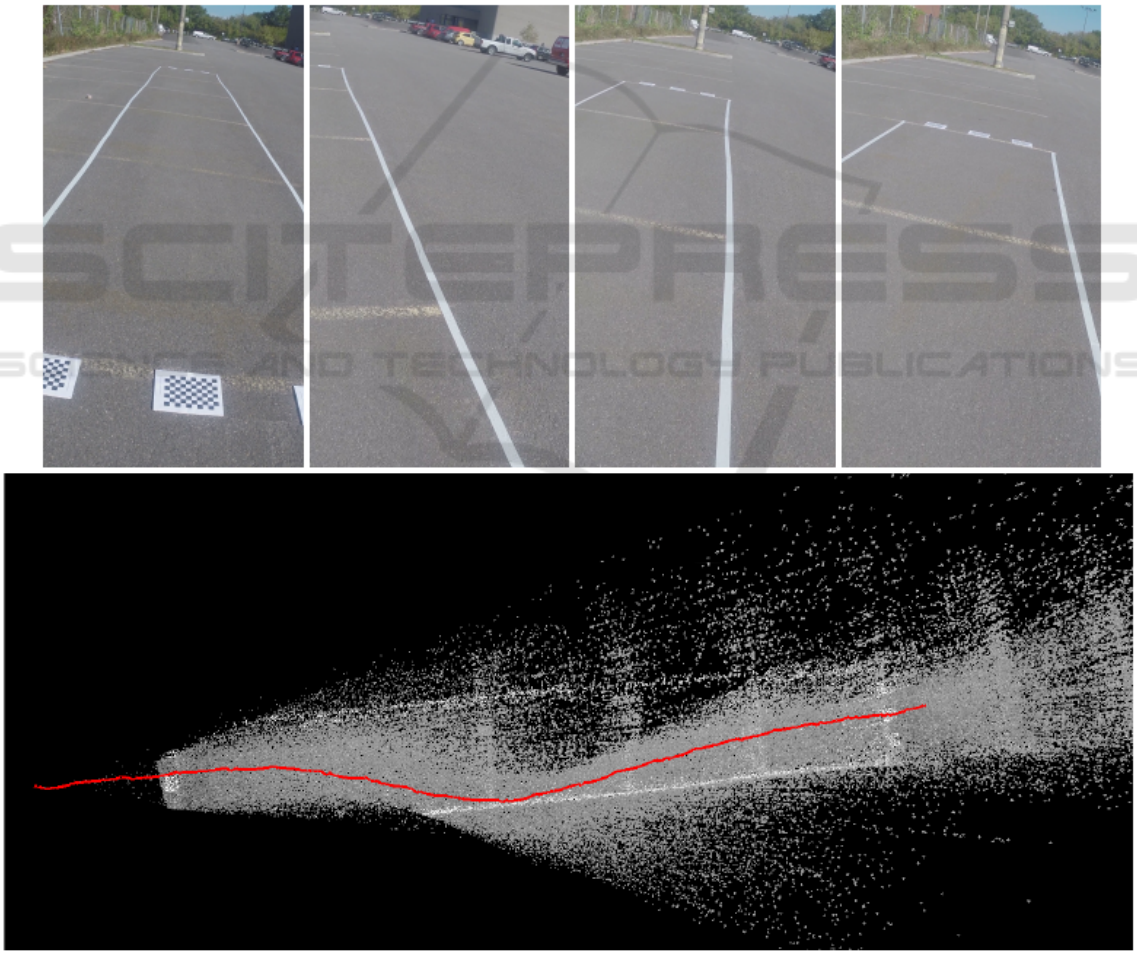

Figure 3: Rendering of a trajectory. Top: video frames at 0s, 9s, 10s and 12s of the 16s video clip. Bottom: top view of the recovered trajectory. The strong deviation in the middle of the sequence is clearly visible.

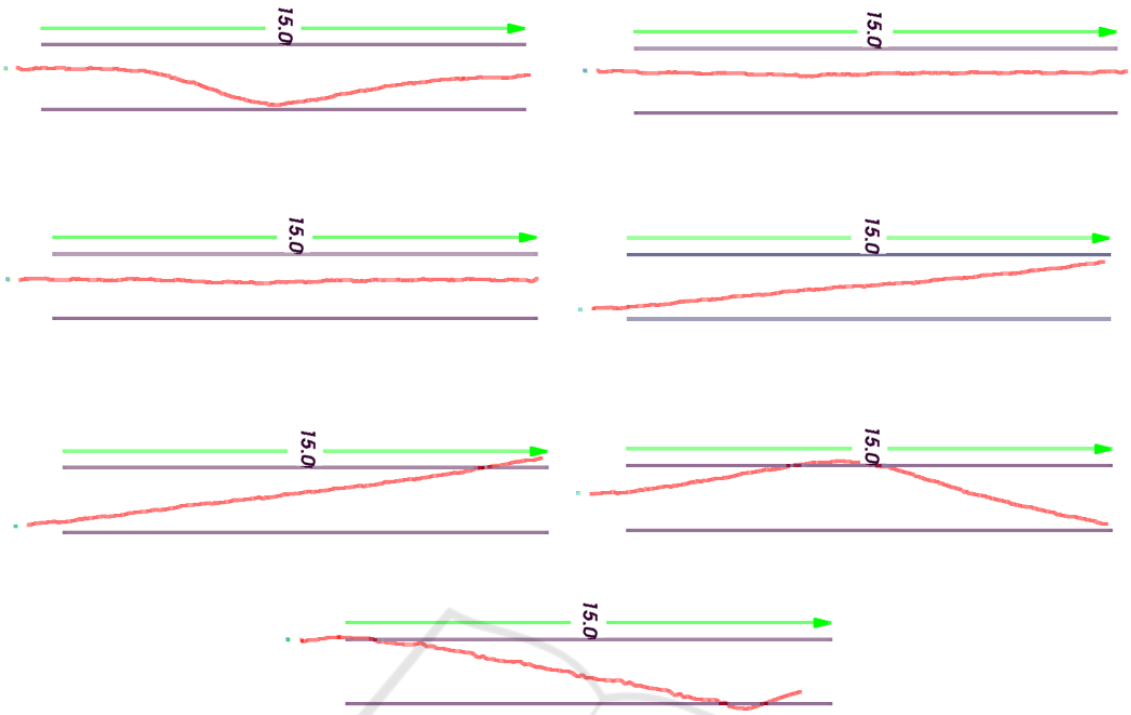

Figure 4: Additional results for seven crossings. Trajectories are shown in red, and the pair of blue lines represents the 15m-long crosswalk.

the camera pose from which the algorithm cannot recover. This observation underlines the importance of carefully designing the experimental protocol when subjects will be involved in a real setting.

Another observation is about the high number of feature points that DSO can track as the camera moves. As opposed to many competing methods, DSO does not search for visual keypoints, but instead splits each video frame into blocks and retains a number of candidate points with high image gradients in each block. The strategy makes sure that points are well distributed throughout the frame, and even across frames. Block size and gradient threshold are also dynamically set to ensure that the pool of candidate points is sufficiently rich for the camera pose estimate. Consequently, this strategy enables DSO to perform well in the current context, even though a weakly textured surface (pavement) occupies a significant portion of the image. Other methods such as ORB-SLAM would have failed to provide any reasonable odometry results in a similar context.

The accuracy of the odometry can be assessed using two sequences where the subject was asked to walk in the center of the corridor (see Fig. 4). Considering that both trajectories are about 10% off the corridor centerline (20cm on the left-hand side) and that the camera held by the subject was 8cm off the body centerline (on the left-hand side as well) for mechanical reasons, a rough evaluation gives an error of about 12cm for these two sequences. These encouraging results justify the planning of a formal evaluation involving blind persons in a real street intersection, which will allow us to collect more accurate performance measures. It will be interesting to evaluate the stability of visual odometry in the presence of cast shadows and moving objects such as pedestrians, bicycles, etc.

5 CONCLUSION

This paper reported on preliminary tests involving visual odometry for localizing people in a street crosswalk. Our objective is to measure the ability of a blind person to engage in a crossing and stay on course by listening to the accessible pedestrian signal. Preliminary tests have shown that analyzing the video footage from a wearable camera attached to a person provides enough information to locate them in a street crosswalk via camera pose estimation. Although the focus of the paper is 3D positioning, visual odometry may also allow for the monitoring of the orientation of the person with respect to the crosswalk, for example capturing hesitations as the crossing progresses.

For our live tests, the use of a wearable Inertial Measurement Unit (IMU) may be considered to further improve the accuracy of the odometry algorithm during post-processing.

ACKNOWLEDGEMENTS

This work has been made possible through funding from the Ministère de l'Économie, de la Science et de l'Innovation du Qubec (MESI) of Gouvernement du Québec.

REFERENCES

Brubaker, M. A., Geiger, A., and Urtasun, R. (2016). Map-based probabilistic visual self-localization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(4):652–665.

Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., and Leonard, J. J. (2016). Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Trans. Robotics*, 32(6):1309–1332.

Engel, J., Koltun, V., and Cremers, D. (2017). Direct sparse odometry. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP(99):1–1.

Engel, J., Schöps, T., and Cremers, D. (2014). LSD-SLAM: Large-scale direct monocular SLAM. In *European Conference on Computer Vision*, pages 834–849.

Gutiérrez-Gómez, D., Puig, L., and Guerrero, J. J. (2012). Full scaled 3D visual odometry from a single wearable omnidirectional camera. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 4276–4281.

Klein, G. and Murray, D. (2007). Parallel tracking and mapping for small AR workspaces. In *International Symposium on Mixed and Augmented Reality*, pages 225–234.

Lalonde, M., Chapdelaine, C., and Foucher, S. (2015). Localizing people in crosswalks with a moving handheld camera: proof of concept. In *Proc. SPIE 9405, Image Processing: Machine Vision Applications VIII*, volume 9405.

Laroche, C., Leroux, T., Giguere, C., and Poirier, P. (2000). Field evaluation of audible traffic signals for blind pedestrians. In *Triennial Congress of the International Ergonomics Association*.

Lynen, S., Achtelik, M. W., Weiss, S., Chli, M., and Siegwart, R. (2013). A robust and modular multi-sensor fusion approach applied to MAV navigation. In *IEEE International Conference on Intelligent Robots and Systems*, pages 3923–3929.

Mur-Artal, R., Montiel, J. M. M., and Tardos, J. D. (2015). ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Trans. Robotics*, 31(5):1147–1163.

Pink, O., Moosmann, F., and Bachmann, A. (2009). Visual features for vehicle localization and ego-motion estimation. In *IEEE Intelligent Vehicles Symposium*, pages 254–260.

Scaramuzza, D., Fraundorfer, F., Pollefeys, M., and Siegwart, R. (2009). Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints. In *International Conference on Computer Vision*, pages 1413–1419.

Se, S., Lowe, D., and Little, J. (2002). Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *Int. J. Robotics Research*, 21(8):735–758.

Song, S., Chandraker, M., and Guest, C. C. (2016). High accuracy monocular SfM and scale correction for autonomous driving. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(4):730–743.