Scaled Monocular SLAM for Walking People

Daniel Gutiérrez-Gómez and J.J. Guerrero
Dpto. Informática e Ingeniería de Sistemas (DIIS) - Inst. Investigación en Ingeniería de Aragón (I3A),
Universidad de Zaragoza
María de Luna 1, Edificio Ada Byron, E-50018 Zaragoza, Spain
{danielgg, josechu.guerrero}@unizar.es

ABSTRACT

In this paper we present a full-scaled real-time monocular SLAM using only a wearable camera. Assuming that the person is walking, the perception of the head oscillatory motion in the initial visual odometry estimate allows for the computation of a dynamic scale factor for static windows of N camera poses. Improving on this method we introduce a consistency test to detect non-walking situations and propose a sliding window approach to reduce the delay in the update of the scaled trajectory. We evaluate our approach experimentally on a unscaled visual odometry estimate obtained with a wearable camera along a path of 886 m. The results show a significant improvement respect to the initial unscaled estimate with a mean relative error of 0.91% over the total trajectory length.

Author Keywords

monocular SLAM, visual odometry, wearable vision.

ACM Classification Keywords

I.2.10 Artificial Intelligence: Vision and Scene Understanding — motion.

General Terms

Algorithms, Experimentation.

INTRODUCTION

The research on wearable cameras for personal aiding has widespread since the pioonering works of Mann [12]. Concerning odometric localisation, Mayol-Cuevas et al. [13], presented a wearable active vision systems which are able to change its orientation and use monocular SLAM (Simultaneous Localisation and Mapping) for self-localisation. In [2] Badino and Kanade propose head-wearable stereo system to estimate structure and motion. Alcantarilla et al. [1] propose a wearable stereo SLAM system which is able to detect moving objects of the scene.

In monocular SLAM the true scale of the map and trajectory estimates is not observable due to the unability to measure real distances in one image. This limitation also affects to the initialisation and parameterisation of new observed landmarks. In spite of this, under a proper parameterisation of the bearing and depth of the landmarks [4], monocular SLAM can be performed succesfully without requiring distance measurements or information about the scale.

However as the the trajectory becomes longer, the scale is likely to drift. Since this drift cannot be observed, it deformates the final estimate beyond a simple scale ambiguity. The scale drift can be reduced if the scale is estimated periodically by some way. Alternatively, Strasdat et al. [16] show that, by introducing a gauge degree of freedom on the camera poses, in a loop closure the scale drift can be corrected without estimating the scale.

In this paper we extend on our previous work [8] to produce an scaled visual odometry from a monocular SLAM with a wearable omnidirectional helmet-camera (Fig. 1). Our main contribution is that, while in [8] the scale was estimated offline and the user was assumed to be always walking, in this paper we propose a method which operates in real-time together with the visual SLAM algorithm and also is able to detect special situations like stairs or stops, which could make the original method fail.

RELATED WORK

In the reviewed literature the problem of the scale estimation for monocular SLAM is addressed either by using additional proprioceptive sensors, like IMUs (Intertial Measurement Unit) and odometers, which provide metric information; or by considering geometric priors or constraints.

Lupton and Sukkarieh [11] make the true scale observable by integrating the visual data and the IMU data within an information filter. Nützi et al. [14] propose the fusion of a initial visual odometry estimate with IMU data in an EKF framework to compute the scaled odometry. Engel et al. [7] estimate the scale factor from a quadricopter onboard IMU and

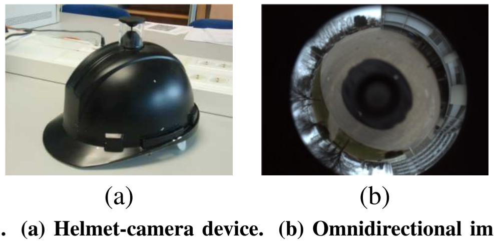

(a) (b)

Figure 1. (a) Helmet-camera device. (b) Omnidirectional image captured with this device.

ISWC '13, Sept 9-12, 2013, Zurich, Switzerland.

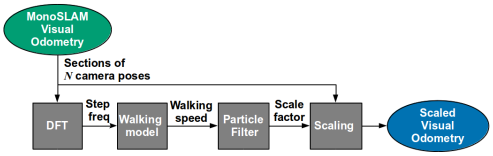


Figure 2. Scheme of the basic scaling method.

altimeter measurements by an optimization scheme. Cumani et al. [5] suggest the combination the wheel odometry and the visual information to obtain the scaled map.

Without using additional sensors, Lothe et al. [10] use the prior knowledge of the distance from the camera to the ground plane to compute the scale factor of the scene. Scaramuzza et al. [15] exploit non-holonomic motion constraints of wheeled vehicles and the knowledge of the horizontal distance between the camera and the rear axis of the vehicle to resolve the scale. Botterill et al. [3] correct the scale drift by identifying previously learnt object classes of the environment.

IMPROVED SCALING APPROACH

To solve the scale problem, we build on our method proposed in [8] which is performed consecutively on sections of the trajectory estimated by a state-of-the-art monocular EKF-SLAM approach adapted to catadioptric cameras in [9]. This method is divided in four steps (Fig. 2). Given a section with N camera poses, first we estimate the step frequency from the vertical oscillatory motion of the head. This is done by computing the Discrete Fourier Transform (DFT) on the z-component of the visual odometry (Fig. 3). Then the step frequency is related to the real walking speed via an empirical model. With the obtained walking speed, the scale factor is estimated dynamically by using a particle filtering approach and then it is applied to the current section. The final output of this method is a full scaled estimate of the visual odometry.

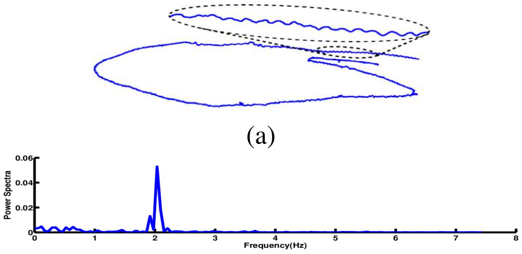
The described scaling method implicitily asumes that the person is always walking and thus may not be robust to frequent situations when the person is not walking like stops, or going up-downstairs. Also, the delay related to the need of taking windows of N frames to compute the DFT yield a low update rate of the scaled estimate in real-time operation. These two problems are addressed in this section.

Detection of non walking situations

To detect non-walking situations we check the spectral power $\bar{P}\left(f_{st,k}\right)$ of the computed step frequency to be consistent with a range of feasible oscillation amplitudes during walking bounded by a A_z^+ and A_z^- . Applying the Parseval's theorem, which states that the energy of a signal is preserved in the frequency domain, we can estimate the energy $\bar{P}$ of the head oscillation signal as:

$$\bar{P}\left(f_{st,k}\right) = 2\frac{F_s}{N}\sum_{m=m^-}^{m^+} \mathit{\Gamma}_d\left(f_{m,k}\right), \qquad (1)$$

with $m^- = \text{round}\left(N\frac{f_{st}-\Delta f}{F_s}\right)$, $m^+ = \text{round}\left(N\frac{f_{st}+\Delta f}{F_s}\right)$, and where $f_{st,k}$ is the estimated step frequency, $\mathit{\Gamma}_d$ is the discretised spectrogram of the z-component of the camera poses, F_s is the sampling frequency of the camera, N the camera poses in the analysed section and Δf the frequency interval centered at $f_{st,k}$ along which the energy is spread along.


(a)

(b)

Figure 3. (a) Trajectory estimate of Visual SLAM from a head-mounted catadioptric camera including a partial zoom. (b) Power spectra of the vertical component

Note that, since the power spectral density is computed for the unscaled z-component of the visual odometry, the computed power must be scaled by multiplying it by the square of the current scale factor estimate $\bar{d}_k$. Thus, assuming that the head oscillation is sinusoidal, the condition for the spectral power consistency of the step frequency yields:

$$\frac{1}{2}{A_z^-}^2 \leq \bar{d}_k^2\bar{P}\left(f_{st,k}\right) \leq \frac{1}{2}{A_z^+}^2. \qquad (2)$$

If this condition is not filled we should choose another strategy. For example if $\bar{d}_k^2\bar{P}\left(f_{st,k}\right) \leq \frac{1}{2}{A_z^-}^2$ we may assume that the person is stopped and then avoid updating the scale factor. Or if $\bar{d}_k^2\bar{P}\left(f_{st,k}\right) \geq \frac{1}{2}{A_z^+}^2$ we may state that the user is on a stairs part and do $V_{walk,k} = Lf_{st,k}$ taking as L an approximation of the distance between two stair steps.

Real time implementation and update delay reduction

The complete scaling algorithm is implemented in a new thread within the real-time monoSLAM C++ application [6], which works in parallel with the main SLAM thread. After each EKF iteration, the main thread stores the last camera poses in a shared buffer. When this buffer is fully updated with N poses, the main thread triggers the scaling thread. After executing the scaling algorithm, the scaled trajectory is updated by adding the recently scaled camera poses.

In spite of the real-time operation the update of the scaled odometry estimate is delayed due to the time it takes to fill the buffer with the N states needed to perform the DFT. One way to reduce this delay is to reduce N, at the expense of reducing also both resolution and accuracy of the DFT to compute the step frequency. Given a low N, padding the data sequence with zeros can improve the resolution but cannot solve the accuracy issues .

Alternatively we propose to use a sliding window, updating only one fraction N_f of the buffer at a time. Thus the number of camera poses used for the spectral analysis remains N by reusing poses from previous sections, while the amount of scaled camera poses per section is reduced to N_f. The complete scaling method is described in Algorithm 1.

Algorithm 1 Complete Visual Odometry Scaling algorithm

```
Require: $C_{k,1..N}$, $S_{k-1}$
Ensure: $\hat{C}_{k,1..N_f}$, $S_k$
  //Notation
  $C_{k,n} =$ $n^{th}$ unscaled camera state of section $k$
  $\hat{C}_{k,n} =$ $n^{th}$ scaled camera state of section $k$
  $N =$ # input camera states
  $N_f =$ # output/new camera states
  $S_k =$ Set of particles for the particle filter
  //Algorithm
  $k = 0$; $[S_0]$ = Initialize particles ()
  while Not end of sequence do
    $k = k + 1$
    Wait for new $C_{k,1..N}$ from monoSLAM
    $\left[z_{k,1..N},\ \mu_{V,k},\ \sigma_{V,k}\right]$ = Extract $z$-comp & mean speed $\left(C_{k,1..N}\right)$
    $\left[z_{k,1..N}\right]$ = High Pass Filter $\left(z_{k,1..N}\right)$
    $\left[f_m,\ \Gamma_{d,k}\right]$ = Spectrogram $\left(z_{k,1..N}\right)$
    $\left[f_{st,k},\ \Gamma_{d,k}\left(f_{st,k}\right)\right]$ = Estimate Step Frequency $\left(f_m,\ \Gamma_{d,k}\right)$
    if Step freq power is consistent $\left(d_{k-1},\ \Gamma_{d,k}\left(f_{st,k}\right)\right)$ then
      $\left[V_{walk,k}\right]$ = Walking speed model $\left(f_{st,k}\right)$
      $\left[S_k\right]$ = Sample Proposal Distribution $\left(S_{k-1},\ \mu_{V,k}, \sigma_{V,k}\right)$
      $\left[S_k\right]$ = Weighting and Resampling $\left(S_k,\ V_{walk,k}\right)$
      $\left[d_k\right]$ = Compute mean scale factor $\left(S_k\right)$
    else
      $d_k = d_{k-1}$ ; $S_k = S_{k-1}$
    end if
    if k=1 then
      $\left[\hat{C}_{1,1..N}\right]$ = Scale Section $\left(d_1,\ C_{1,1..N}\right)$
    else
      $\left[\hat{C}_{k,1..N_f}\right]$ = Scale Section $\left(d_k,\ C_{k,(N-N_f+1)..N}\right)$
    end if
  end while
```

EXPERIMENTS

We acquired an outdoor dataset with a catadioptric omnidirectional camera attached to a helmet. The used camera is the Sentech UltraSmall STC-MC83USB model. The sequence was acquired at a frame rate of 15 fps along a path of 886 m and contains a variety of situations like changes of pace, stops, stairs and walking along a narrow corridor. Transitions between these situations have been ticked accordingly in the frame when they take place. An example of the operation of our scaling approach is provided in the supplementary video file [1].

The Ground Truth has been obtained from the Google Maps satellite view using the distance measurement tool. To compare numerically the Ground Truth and the scaled estimates, we parameterise both curves by a scalar α defined by the normalised accumulated distance which spans from 0 (start) to 1 (end). Then, given the Ground Truth trajectory $\mathbf{t}_{GT}$, to compute the error for a given pose $\mathbf{t}_{VO}^{(i)}$ of the visual odometry we find the point $\mathbf{t}_{GT}^{(j)}$ in the Ground Truth whose normalised arc length is closest to the one of $\mathbf{t}_{VO}^{(i)}$, that is:

$$\mathbf{t}_{GT}^{(j)} = \underset{\mathbf{t}_{GT}}{\arg\min} \|\alpha(\mathbf{t}_{VO}^{(i)}) - \alpha(\mathbf{t}_{GT})\|, \tag{3}$$

$$err^{(i)} = \|\mathbf{t}_{VO}^{(i)} - \mathbf{t}_{GT}^{(j)}\|. \tag{4}$$

Results

We have tested 3 different alternatives for the setup of the algorithm. In two alternatives N_f is set to 50 and N is varied to compare the performance of dynamic ($N = 200$) and static ($N = 50$) windows. The third alternative is a static window with $N = 25$. Lower values for N cannot be considered since step frequency cannot be measured due to the lack of enough camera poses to capture the head oscillations. To have the same resolution in the DFT, the data sequence is zero padded to $N_p = 256$ in the three considered cases.

Fig. 4 (top) reveals how the step frequency computed from the raw unscaled visual odometry varies accordingly with the pace of the walker. Note how for lower values of N the step frequency estimate along the trajectory is less accurate and oscillates more, though the global tendence in the pace is still captured. In Fig. 4 (bottom), during the long stop, the spectral power shows a violation of the consistency condition which leads to ignore the erratic estimation of the step frequency. During the short stop, the violation of the consistency is not observed with $N = 200$, due to the masking effect of the poses corresponding to a walking state. In the case of going upstairs the magnitude of the power peak is too low to establish a clear upper limit for the consistency condition.

In Fig. 5 (top) it can be observed how the scale factor is dinamically estimated. The major amount of drift in the scale occurs at two parts of the trajectory. The first one occurs moments before the long stop (around frame 5500) and consists of a sharp variation due to a sudden loss of most of the tracked features (maybe due to a image matching perturbation or a lost frame). The second one occurs gradually along a narrow

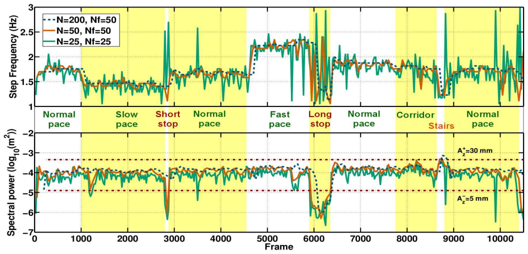


Figure 4. Evolution of the step frequency estimate (top) and its corresponding spectral power (bottom). Consistency bounds are violated when estimate does not corespond to a walking step frequency.

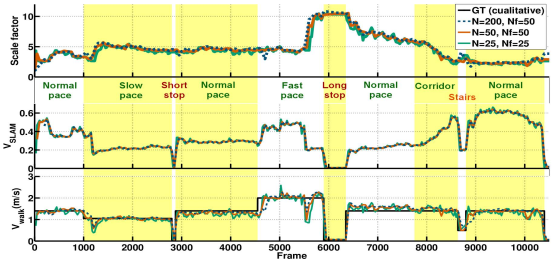


Figure 5. Evolution of (top) the scale factor, (middle) the non-dimensional speed and (bottom) the real walking speed which results of multiplying the scale factor by the non-dimensional speed. The comparison with a cualitative Ground Truth showing the change in pace gives an idea of the accuracy of the method.

[1] `http://www.youtube.com/watch?v=fcQ72SgqeZM`

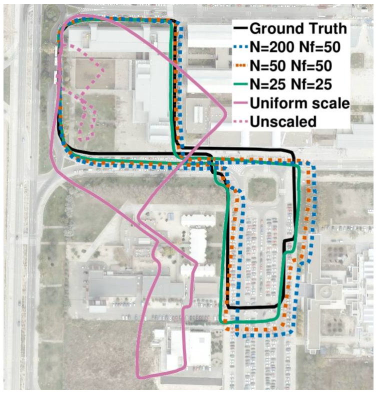


Figure 6. Scaled and unscaled trajectories compared with the Ground Truth.

N	N_f	Mean error[m]	Maximum error[m]	Relative mean error
200	50	14.75	26.80	1.66%
50	50	11.54	20.34	1.30%
25	25	8.05	15.44	0.91%
Uniform scale		69.97	146.45	7.90%

Table 1. Estimation error for combinations of N and N_f.

corridor between two buildings on frames from 7800 to 8600, where the scale is lower compared to an wider environment.

Finally in Fig. 6 we show a comparison between the scaled trajectories obtained for each considered setup and both the unscaled and the uniformly scaled raw visual odometries. Loops have been closed at the begining and in the middle of the path, by manually introducing the constraints and performing an optimisation. The great improvement of all the considered cases respect to the uniformly scaled visual odometry show the ability of our approach to remove the scale drift. This drift removal, though not complete, allows to obtain a close approximation to the Ground Truth.

Numerical comparison with the Ground Truth is detailed in table 1. It is shown that taking $N = 25$ offers an slightly better scaled estimate of the visual odometry with a mean relative error of 0.91% over the trajectory length. Using a sliding window approach taking a higher N we may expect a better estimate since we use more data to compute the DFT and thus, as shown in Fig. 4, we obtain a more accurate and stable step frequency. However, it must be noted first, that it is achieved at the expense of delaying the detection of transitions in the walking pace and even not observing fast events like the short stop; and secondly, that a sliding window approach computes a scale factor taking measurements obtained from different data: while we take N camera poses to compute V_{walk}, we only take the last N_f to compute V_{SLAM}. This can be a source of inaccuracies when transitions occur.

CONCLUSIONS

In this paper we have presented a scaling algorithm for a monocular SLAM with a wearable omnidirectional camera. The main contributions of this paper are two: the detection of non-walking situations; and a evaluation of a sliding window to reduce the delay in the update of the scaled trajectory. Experimental results in a path of 886 m show that our approach is able to accurately compute the scale with drift correction along a path under a variety of situations needing no more data than the visual odometry itself. Although the results using different setups are similar, the use of a sliding window is discouraged in favour of narrower static windows, since they reduce the update time and are able to capture short changes in the walking pattern.

ACKNOWLEDGEMENTS

This work was supported by project DPI2012-31781 and scholarship FPU12-05507.

REFERENCES

1. Alcantarilla, P., Yebes, J., Almazán, J., and Bergasa, L. On combining visual slam and dense scene flow to increase the robustness of localization and mapping in dynamic environments. In *IEEE Int. Conf. on Robotics and Automation (ICRA)* (2012).
2. Badino, H., and Kanade, T. A head-wearable short-baseline stereo system for the simultaneous estimation of structure and motion. In *Proc. IAPR Conf. on Machine Vision Applications (MVA)* (2011).
3. Botterill, T., Mills, S., and Green, R. Correcting scale drift by object recognition in single camera slam. *IEEE Trans. on Systems, Man, and Cybernetics–Part B: Cybernetics* (2012).
4. Civera, J., Davison, A. J., and Montiel, J. M. M. Inverse depth parametrization for monocular slam. *IEEE Trans. on Robotics 24*, 5 (2008), 932–945.
5. Cumani, S., Denasi, A., Guiducci, A., and Quaglia, G. Integrating monocular vision and odometry for slam. in wseas transactions on computers. *WSEAS Trans. on Computers 3* (2004), 625630.
6. Davison, A. J., Reid, I. D., Molton, N. D., and Stasse, O. Monoslam: Real-time single camera slam. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 29* (2007), 1052–1067.
7. Engel, J., Sturm, J., and Cremers, D. Camera-based navigation of a low-cost quadrocopter. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robot Systems (IROS)* (2012).
8. Gutierrez, D., Puig, L., and Guerrero, J. J. Full scaled 3d visual odometry from a single wearable omnidirectional camera. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robot Systems (IROS)* (2012).
9. Gutiérrez-Gómez, D., Rituerto, A., Montiel, J. M. M., and Guerrero, J. J. Adapting a real-time monocular visual slam from conventional to omnidirectional cameras. In *11th OMNIVIS, held with International Conference on Computer Vision (ICCV)* (2011).
10. Lothe, P., Bourgeois, S., Royer, E., Dhome, M., and Naudet-Collette, S. Real-time vehicle global localisation with a single camera in dense urban areas: Exploitation of coarse 3d city models. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (2010), 863–870.
11. Lupton, T., and Sukkarieh, S. Removing scale biases and ambiguity from 6dof monocular slam using inertial. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)* (2008), 3698–3703.
12. Mann, S. Wearable computing: A first step toward personal imaging. *Computer 30*, 2 (Feb. 1997), 25–32.
13. Mayol, W. W., Davison, A. J., Tordoff, B. J., and Murray, D. W. Applying active vision and slam to wearables. In *Proc. Int. Symp. on Robotics Research (ISRR)* (2003), 325–334.
14. Nützi, G., Weiss, S., Scaramuzza, D., and Siegwart, R. Fusion of imu and vision for absolute scale estimation in monocular slam. *J. of Intelligent Robotic Systems 61*, 1-4 (2010), 287–299.
15. Scaramuzza, D., Fraundorfer, F., Pollefeys, M., and Siegwart, R. Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints. In *Proc. IEEE Int. Conf. on Computer Vision* (2009), 1413–1419.
16. Strasdat, H., Montiel, J. M. M., and Davison, A. Scale drift-aware large scale monocular slam. In *Proc. Robotics: Science and Systems (RSS)* (2010).