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 pionering 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 inability 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 deforms 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

Figure 1. (a) Helmet-camera device. (b) Omnidirectional image captured with this device.
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 of this method by using a particle filtering approach.

The described scaling method implicitly asumes that the percentage 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 of this method by using a particle filtering approach.

**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 implicitly assumes 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 \( \bar{P} (f_{st,k}) \) 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:

\[
\bar{P} (f_{st,k}) = 2 \frac{F_s}{N} \sum_{m=\pm \frac{N}{2}}^{\frac{N}{2}-1} \Gamma_d (f_{m,k}),
\]

with \( m = \text{round} (N \frac{f_{st,k}-\Delta f}{F_s}) \), \( m = \text{round} (N \frac{f_{st,k}+\Delta f}{F_s}) \), and where \( f_{st,k} \) is the estimated step frequency, \( \Gamma_d \) is the discrete-

\[ \frac{1}{2} A_z^{+2} \leq \bar{A}_z (f_{st,k}) \leq \frac{1}{2} A_z^{-2}. \]

If this condition is not filled we should choose another strategy. For example if \( \bar{A}_z (f_{st,k}) \leq \frac{1}{2} A_z^{+2} \) we may assume that the person is stopped and then avoid updating the scale factor. Or if \( \bar{A}_z (f_{st,k}) \geq \frac{1}{2} A_z^{-2} \) we may state that the user is on a stairs part and do \( V_{walk,k} = L f_{st,k} \) 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: \( C_{k,1..N_f}, S_k \)

Ensure: \( C_{k,1..N_f}, S_k \)

//Notation
- \( C_{k,n} \) = \( n^{th} \) unscaled camera state of section \( k \)
- \( C_{k,n} \) = \( k^{th} \) 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 ()]
- \( \text{while Not end of sequence do} \)
  - \( k = k + 1 \)
  - Wait for new \( C_{k,1..N} \) from monoSLAM
  - \([z_{k,1..N}, \mu_{V,k}, \sigma_{V,k}] = \text{Extract z-comp & mean speed} \ (C_{k,1..N}) \]
  - \([\hat{z}_{k,1..N}] = \text{High Pass Filter} \ (z_{k,1..N}) \]
  - \([f_{m}, \Gamma_{d,k}] = \text{Spectrogram} \ (z_{k,1..N}) \]
  - \([f_{st,k}, \Gamma_{d,k} (f_{st,k})] = \text{Estimate Step Frequency} \ (f_{m}, \Gamma_{d,k}) \]
  - If Step freq power is consistent (\( d_{k-1}, \Gamma_{d,k} (f_{st,k}) \)) then
    - \([V_{walk,k}] = \text{Walking speed model} \ (f_{st,k}) \]
    - \([S_k] = \text{Sample Proposal Distribution} \ (S_{k-1}, \mu_{V,k}, \sigma_{V,k}) \]
    - \([S_k] = \text{Weighting and Resampling} \ (S_k, V_{walk,k}) \]
    - \([d_k] = \text{Compute mean scale factor} \ (S_k) \]
  - else
    - \( d_k = d_{k-1}; S_k = S_{k-1} \)
- \( \text{end if} \)
- \( \text{if } k = 1 \) then
  - \([\hat{C}_{1,1..N}] = \text{Scale Section} \ (d_{1}, C_{1,1..N}) \]
- \( \text{else} \)
  - \([\hat{C}_{k,1..N_f}] = \text{Scale Section} \ (d_k, C_{k,(N-N_f+1)..N}) \]
- \( \text{end if} \)
- \( \text{end while} \)

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. 

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 \( \alpha \) defined by the normalised accumulated distance which spans from 0 (start) to 1 (end). Then, given the Ground Truth trajectory \( t_{GT} \), to compute the error for a given pose \( t_{VO}^{(i)} \) of the visual odometry we find the point \( t_{GT}^{(i)} \) in the Ground Truth whose normalised arc length is closest to the one of \( t_{VO}^{(i)} \), that is:

\[
t_{GT}^{(i)} = \arg \min_{t_{GT}} \| \alpha(t_{VO}^{(i)}) - \alpha(t_{GT}) \|, \tag{3}
\]

\[
e_{err}^{(i)} = \| t_{VO}^{(i)} - t_{GT}^{(i)} \|. \tag{4}
\]

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 tendency 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 a violation 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 dynamically 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 correspond 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 qualitative Ground Truth showing the change in pace gives an idea of the accuracy of the method.

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

Table 1. Estimation error for combinations of \( N \) and \( N_f \).

<table>
<thead>
<tr>
<th>( N )</th>
<th>( N_f )</th>
<th>Mean error[m]</th>
<th>Maximum error[m]</th>
<th>Relative mean error</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>50</td>
<td>14.75</td>
<td>26.80</td>
<td>1.06%</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>11.54</td>
<td>20.54</td>
<td>1.30%</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>8.05</td>
<td>15.44</td>
<td>0.91%</td>
</tr>
<tr>
<td>Uniform scale</td>
<td>69.97</td>
<td>146.45</td>
<td>7.90%</td>
<td></td>
</tr>
</tbody>
</table>

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