Camera based step detection on mobile phones

Ferenc Aubeck*, Carsten Isert* and Dominik Gusenbauer*

* BMW Group Research and Technology, Hanauer Straße 46, D-80992 Munich, forename.surname@bmw.de

***Abstract*—This paper presents an extension to an indoor positioning system leveraging the camera which is built into current mobile phones to detect steps. Current indoor positioning systems on mobile phones based on pedestrian dead reckoning (PDR) often rely on step length and step frequency estimation. Usually the accelerometer is used to determine these values. Depending on the attitude of the device or the smoothness of the movement sometimes steps are not correctly detected. We propose to use the camera as additional sensor element to detect the forward section of both feet appear and disappear in the camera image and detect the steps based on this information.**

***Keywords*—pedestrian dead reckoning, camera, pedometer, step estimation, step frequency, indoor navigation**

I. Introduction

To offer a continuous navigation experience, an indoor navigation system must be able to determine a pedestrian's trajectory in both presence and absence of absolute positioning infrastructure like GPS, WLAN or others. The availability of a wide range of sensors in current smart phones like the iPhone 4 or the new generation of Android based devices enables software solutions for inertial positioning. These sensors include gyroscopes, acceleration sensors and magnetometers. The examination of a pedestrians acceleration signal is a common strategy to analyze human locomotion, detect steps and estimate each steps length [1][2]. Thus PDR is well suited for the present application, providing independent means of position determination.

So far, cameras have also been used for positioning in the context of mobile phones, but mostly for scanning QR codes or other preinstalled markers and derive absolute positions from this information [3]. Other examples of fusing inertial measurement and vision based systems are mainly focused on improving the calibration of the camera itself [4], estimating the camera's ego motion [5] or improving real-time camera pose tracking in the context of augmented reality [6][7].

We propose to use the camera as an additional sensor in PDR solutions. Using image processing the appearance and disappearance of the user's feet can be detected and used as additional signal for step counting in an application scenario where the user is holding the device in front of him.

II. PDR System Architecture

Like in many other studies, our pedestrian dead reckoning algorithm is based on the examination of the pedestrians acceleration signal. It therefore shares the same difficulties to correctly detect steps, to estimate each step's length according to a current movement status and to determine a correct heading.

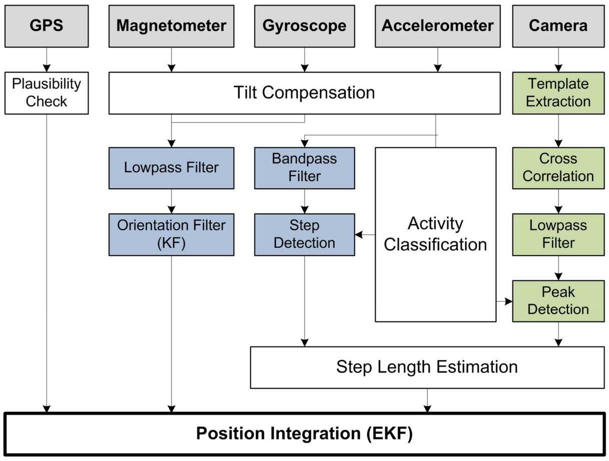

Figure 1. The basic system architecture

The basic system architecture is initially based on the solution presented in [8], but provides further improvements by incorporating an additional tri-axial gyroscope motion sensor and a camera as depicted in Fig. 1.

As our strategy is to implement our solutions on off-the-shelf components we are currently developing our prototype on the iPhone 4 because of the newly available gyroscope sensor of the iPhone 4. The gyroscope is used to supplement the electronic magnetic compass and improve the reliability of heading determination. Since this type of sensor is unsusceptible to environmental inferences, the coupling with a magnetic compass provides superior information compared to the one obtained from each module separately and therefore is a common strategy to determine the walking line of sight [9].

To improve the results of the accelerometer based step counting algorithm we are using the device's built-in camera. We capture the movements of the user's feet to determine individual steps based on morphological image processing. By having an additional and user acceleration independent source for step detection and step frequency estimation respectively, the probability of miss-detected steps due to user generated irregular movements such as shaking or periodically swinging the device while walking can be reduced. Another problem for accelerometer based step detection is very slow movement or when the device is held smoothly.

In fact an approach like this depends on the absolute attitude of the device, which is assumed to be located in the user's hand, in front of the body oriented towards the

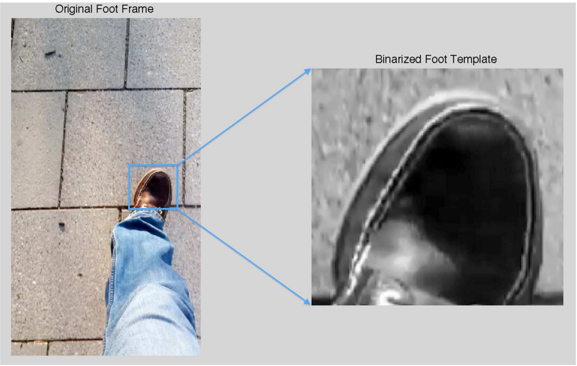

Figure 2. One frame (left) and the saved image template (right) from the right foot.

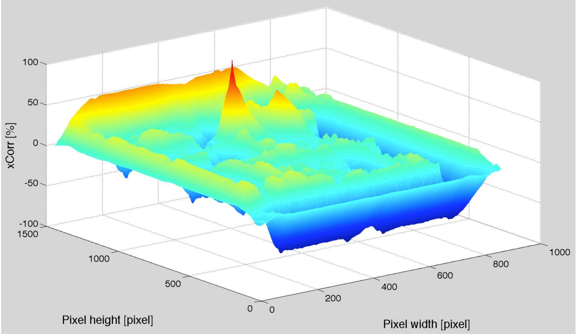

Figure 3. The peak in the image area of the cross-correlation matrix occurs where the template best correlates.

current line of sight with the user supposed to be looking at a map of the environment.

We conducted a user study with 10 persons of different age, height, handedness and gender walking in indoor and outdoor environments holding an iPhone 4. Even with some users not aware of the intended function with every person the feet were generally visible so that our camera based detection algorithm could work. Therefore we conclude that it seems to be a very natural way to hold the phone and our approach will work in a wide range of application areas.

However, it must be mentioned that in some cases the geometrical architecture of the phone can have an influence in such a way that the camera is hidden by the user's fingers, light conditions are not suitable for image processing, or when walking around corners sometimes one foot was not visible so that accelerometer based step detection is still needed.

III. Camera Based Step Estimation

A. *Cross-correlation based template matching algorithm*

The general architecture of our solution is based on standard methods known in image processing and follows a template-based approach [10]. In a first step a sub-image of a single frame is extracted from a standalone image (see Fig. 2). For each foot a template is stored as a matrix. The template for the second foot is mirrored from the template of the first foot.

The camera is used for capturing grayscale image frames. These frames are used in combination with the image templates of the feet to calculate the normalized cross-correlation. This method is independent of differences of brightness and contrast due to the normalization. The correlation maximum value is used to determine the foot motion in the spatial area (Fig. 3).

The normalized cross-correlation [11] is a standard method for matching a template in a given frame by using the following procedure:

1. Calculate the position and correlation between the actual frame *f* and template *t* in the classical two-dimensional spatial area according to (1). *f(x,y)* denotes the intensity value of the frame *f*. *c(u,v)* is the maximum value position and is computed for each point *(u,v)* in the frame *f* and the template *t* by shifting *u* steps in the *x* direction and *v* steps in the *y* direction. The maximum value *c(u,v)* is dependent on the size of the template *t*.

$$c(u,v)=\sum_{x,y} f(x,y)t(x-u,y-v) \tag{1}$$

Equation (1) is sensitive to changes in the frame values (lighting conditions). Instead of maximum value cross-correlation, the cross-correlation coefficient *γ(u,v)* overcomes these difficulties by normalizing the image frame *f* and template *t* vectors to normalized unit length. In (2) vector $\bar{t}$ is the mean of the template *t(x,y)* and the vector $\bar{f}_{u,v}$ is the mean of *f(x,y)* in the template *t* region. Equation (2) is calculating the local sums *(x,y)* over the image frame function *f(x,y)* by precomputing the squared image frame function $f^2(x,y)$ (running sums), once for each frame *f* and at each position *(u,v)*, at which the normalized cross-correlation coefficient is evaluated.

$$\gamma(u,v)=\frac{\sum_{x,y}\left[f(x,y)-\bar{f}_{u,v}\right]\left[t(x-u,y-v)-\bar{t}\right]}{\left\{\sum_{x,y}\left[f(x,y)-\bar{f}_{u,v}\right]^2\sum_{x,y}\left[t(x-u,y-v)-\bar{t}^{\,2}\right]\right\}^{0,5}} \tag{2}$$

2. When using the normalized cross-correlation coefficient value over a frame the number of calculations does no longer depend on the size of the template, but on the size of the image frame.

B. *Step estimation*

The cross-correlation signals c_{left} and c_{right} for the left and right foot are summarized in each frame to restrict the evaluation on one signal according to the formula:

$$c_{sum}=\sqrt{\left(c_{left}\right)^2+\left(c_{right}\right)^2} \tag{3}$$

The obtained signals need to be filtered to reduce the noise and to ensure that the signals are independent from short-term fluctuations. We are using an equiripple band pass filter (Fig. 4).

For filter initialization we used a delay of 20 frames which corresponds approximately to one slow step motion. When a peak is detected over a given sampling sequence, this is identified as a step.

These results can be used as input for the pedometer unit and fused with a step counter and step length estimation element.

The actual normalized cross-correlation coefficient value must be above the maximum noise value in the

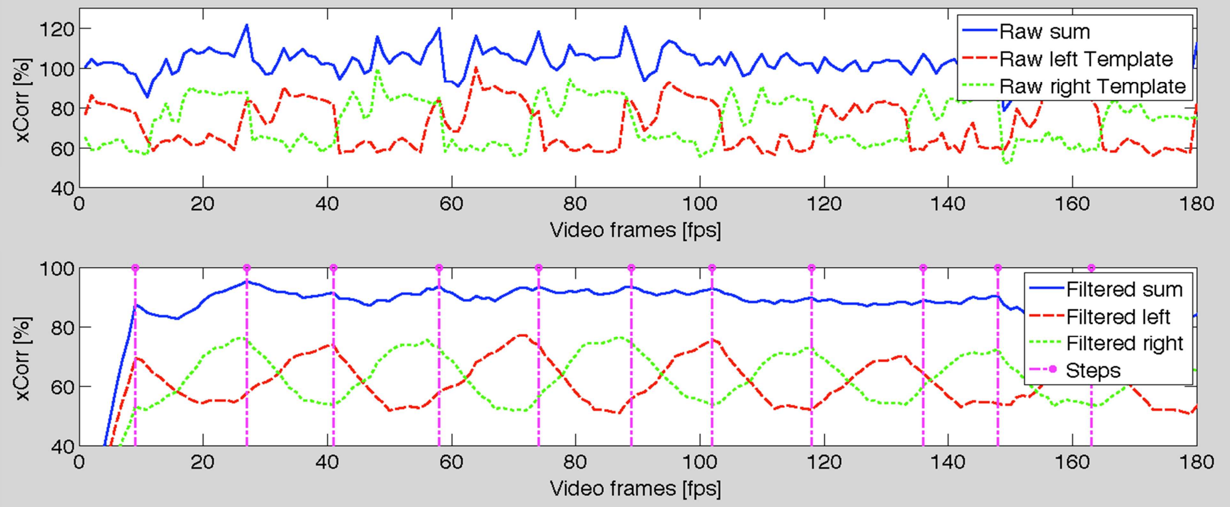

Figure 4. Upper: Raw left-, right- and summarised- foot cross-corelation signals. Bottom: Bandpass filtered signals. The template cross-correlation coefficient over 180 image frames. The foot motion over 11 steps.

actual frame's spatial area to enable the step-detection and to recognize the feet.

C. Automatic template generation

Currently the template matrix is generated by manual selection. The right and left foot template is mirrored from itself depending on the selected side. For user friendly usage the template generation must be done automatically and should also be able to dynamically adapt to different lighting conditions and form of the shoes and the environment.

Our results have shown that the recognition of the feet is very stable and works also when the image or template quality is very poor. The size and movement pattern of the template is also very similar for most use cases. Therefore we propose to solve the problem of template generation by starting with a default template with a general form of a shoe and dynamically adapting the template to the current conditions when the correlation drops below a threshold or the template can be determined from the last stride. The template should be saved when the application is terminated and recovered at startup.

IV. Results and Conclusion

We evaluated the general feasibility of our idea with videos taken with an iPhone 4 with a resolution of 360 x 480 pixels and a frame rate of 24 fps or more. The size of the template was about 50 x 50 pixels. The image processing was done using Matlab Image Processing Toolbox. The average computation time to extract the steps was about 2 minutes for 1 minute of video without any optimizations. We are currently working on implementing this solution directly on the iPhone and integrating it into our indoor positioning framework. Performance optimizations like the reduction of the frame rate, the increase of the image frame resolution through decimation, the restriction of the computation to special areas and the adaption to the hardware are necessary steps.

Our empirical results have shown that for the 10 users in our simple test scenario the conditions enable the usage of the camera as a sensor for step frequency detection and all appearing steps could be detected. To enlarge the optical coverage area the image processing could be based on the full size of the photo and not restricted to the video size.

The fusion of camera and accelerometer inputs seems a promising approach. As an alternative, this algorithm can also be used in applications that already use the camera for marker based positioning to enable dead reckoning between these markers or in augmented reality applications.

Future work will include evaluation of larger distances and the comparison to accelerometer based step detection algorithms. Additionally, we plan to investigate the potential of motion detection algorithms for direct distance and velocity measurements based on the camera image.

References

[1] Q. Ladetto, "On Foot Navigation: Continuous Step Calibration using both Complementary Recursive Prediction and adaptive Kalman Filtering," in Proc. of the 13th Int. Technical Meeting of the Satellite Division of the Institute of Navigation (ION GPS'00), pp. 1735–1740, 2000.

[2] V. Gabaglio, "GPS/INS Integration for Pedestrian Navigation," PhD Thesis, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Publ. No. 2704, 2002.

[3] A. Mulloni, D. Wagner, D. Schmalsteig, and I. Barakonyi, "Indoor Positioning and Navigation with Camera Phones," IEEE Pervasive Computing, vol. 8, no. 2, pp. 22–31, April 2009.

[4] D. Randeniya, M. Gunaratne, S. Sarkar, and A. Nazef, "Calibration of inertial and vision systems as a prelude to multi-sensor fusion," in Transportation Research Part C (Emerging Technologies), vol. 16/2, pp 255–274, 2008.

[5] D. Aufderheide, W. Krybus, "Towards Real-Time Camera Egomotion Estimation and Three-Dimensional Scene Acquisition from Monocular Image Streams," in Proc. of the Int. Conference on Indoor Positioning and Indoor Navigation (IPIN), pp. 13–22, 2010.

[6] J. Hol, T. Schön, F. Gustafsson, and P. Slycke, "Sensor Fusion for Augmented Reality," in 9th Int. Conference on Information Fusion, Florence, Italy, 2006.

[7] G. Bleser, and D. Stricker, "Advanced tracking through efficient image processing and visual-inertial sensor fusion," in Proc. of IEEE Virtual Reality Conference 2008 (VR'08), pp. 137–144, 2008.

[8] D. Gusenbauer, C. Isert, J. Krösche, "Self Contained Indoor Positioning on Off-The-Shelf Mobile Devices," in Proc. of the Int. Conference on Indoor Positioning and Indoor Navigation (IPIN), pp. 834–842, 2010.

[9] J. Collin, “Investigations of Self-Contained Sensors for Personal Navigation,” PhD Thesis, Tampere University of Technology, Tampere, Publ. No. 619, 2006.

[10] G.A. Baxes, Digital Image Processing: Principles and Applications. John Wiley and Sons, pp. 86-99, 1994.

[11] Lewis JP. Fast Normalized Cross-Correlation. Industrial Light and Magic, pp. 1–7, 1996.