MOTION VECTOR ESTIMATION OF A STEREOVISION CAMERA WITH INERTIAL SENSORS

Paweł Pełczyński, Bartosz Ostrowski, Dariusz Rzeszotarski

Technical University of Łódź, Institute of Electronics, Wólczańska 211/215, 90-924 Łódź, Poland (✉ pawel.pelczynski@p.lodz.pl, +48 42 631 26 17, bartosz.ostrowski@p.lodz.pl, dariusz.rzeszotarski@p.lodz.pl)

Abstract

The aim of the presented work was the development of a tracking algorithm for a stereoscopic camera setup equipped with an additional inertial sensor. The input of the algorithm consists of the image sequence, angular velocity and linear acceleration vectors measured by the inertial sensor. The main assumption of the project was fusion of data streams from both sources to obtain more accurate ego-motion estimation. An electronic module for recording the inertial sensor data was built. Inertial measurements allowed a coarse estimation of the image motion field that has reduced its search range by standard image-based methods. Continuous tracking of the camera motion has been achieved (including moments of image information loss). Results of the presented study are being implemented in a currently developed obstacle avoidance system for visually impaired pedestrians.

Keywords: passive navigation, motion vector, disparity, feature points.

1. Introduction

There are a lot of technical fields that utilize automatic navigation using video-based scene analysis systems, such as creating maps based on aerial and satellite imagery, robotics or mobile systems for synthetic presentation of information in the context of the location in the scene (so-called augmented reality systems). An example of such an application is the auditory obstacle presentation system for blind pedestrians under development at the Technical University of Łódź.

We can distinguish between global navigation, which consists in determining the geographical position, and local navigation. Local navigation is a determination of one's position relative to the local reference system associated with the static scene (called navigation frame, *n-frame* [1]) or estimation of scene motion in the reference system associated with the mobile system realizing the task of navigation (called body frame, *b-frame*). Estimation of scene motion in the *b-frame* is used, among other things, for tracking objects in image sequences, which is the main objective of the developed scene analysis system.

2. Local navigation on the basis of a stereoscopic image sequence

A number of basic dependencies and assumptions have been made to allow the development and verification of the algorithm for motion estimation of the stereoscopic camera. The right camera is placed at the origin of the *b-frame* reference, denoted further by index *C*. The location of the camera in the *n-frame* is described by a rigid body motion equation [2]. The coordinates of any point $[X,Y,Z]^T_W$ from the surrounding three-dimensional scene, in the local *n-frame* reference associated with the Earth's surface, and designated by

index *W*, can be related by the following formula to its coordinates in the camera frame of reference *C*:

$$
\begin{bmatrix} X_W \\ Y_W \\ Z_W \end{bmatrix} = \mathbf{R}_{WC} \begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} + \mathbf{T}_{WC} = \begin{bmatrix} \mathbf{r}_1 \ \mathbf{r}_2 \ \mathbf{r}_3 \end{bmatrix} \begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} + \begin{bmatrix} T_{XC} \\ T_{YC} \\ T_{ZC} \end{bmatrix} = \begin{bmatrix} r_{11} \ r_{12} \ r_{13} \\ r_{21} \ r_{22} \ r_{23} \\ r_{31} \ r_{32} \ r_{33} \end{bmatrix} \begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} + \begin{bmatrix} T_{XC} \\ T_{YC} \\ T_{ZC} \end{bmatrix}. \tag{1}
$$

$\mathbf{R}_{WC}$ is the rotation matrix of the camera, $\mathbf{T}_{WC}$ is the translation vector. The above dependence allows to determine the relative displacement of the cameras in the scene based on the estimation of displacements of static objects in the *C* reference associated with the camera. A new position of the point from the surrounding static scene (resulting from the motion of the video system) can be described by the equation:

$$
\begin{bmatrix} X_{CK} \\ Y_{CK} \\ Z_{CK} \end{bmatrix} = \mathbf{R}_C \begin{bmatrix} X_{CP} \\ Y_{CP} \\ Z_{CP} \end{bmatrix} + \mathbf{T}_C. \tag{2}
$$

$\mathbf{R}_C$ is the rotation matrix of the scene, $\mathbf{T}_C$ is the translation vector of the scene, the index *C* indicates the reference system associated with the camera, *P* - initial position, *K* – final position. Further dependences are described in the three-dimensional reference system associated with the initial position of the camera and the index *C* is omitted. Let the position of *P* and K be related to consecutive images of a sequence. Assuming that the change of position and orientation of the stereovision system is small between *P* and K, equation (2) can be approximated by the relation:

$$
\begin{bmatrix} X_K \\ Y_K \\ Z_K \end{bmatrix} = \begin{bmatrix} 1 & \gamma & -\beta \\ -\gamma & 1 & \alpha \\ \beta & -\alpha & 1 \end{bmatrix} \begin{bmatrix} X_P \\ Y_P \\ Z_P \end{bmatrix} - \begin{bmatrix} U \\ V \\ W \end{bmatrix}. \tag{3}
$$

α, β i γ are the angles of the camera rotation around the axes *X, Y* and *Z*, $[U \ V \ W]^T$ is the translation vector of the camera in the scene (Fig. 1). The minus sign before the translational component means that it is the result of camera movement (egomotion [3]) in a static scene, not the scene itself. The increment of coordinates of a point between successive images is equal:

$$
\begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix} = \begin{bmatrix} X_K \\ Y_K \\ Z_K \end{bmatrix} - \begin{bmatrix} X_P \\ Y_P \\ Z_P \end{bmatrix} = \begin{bmatrix} 0 & \gamma & -\beta \\ -\gamma & 0 & \alpha \\ \beta & -\alpha & 0 \end{bmatrix} \begin{bmatrix} X_P \\ Y_P \\ Z_P \end{bmatrix} - \begin{bmatrix} U \\ V \\ W \end{bmatrix}. \tag{4}
$$

After we apply the projective transformation for the pinhole camera, coordinates (*x,y*) of the scene point image are obtained:

$$
x = \frac{fX}{Z}; y = \frac{fY}{Z}, \tag{5}
$$

where *f* is the focal length of the camera.
 If the restriction is imposed on the minimum distance of the camera from objects in the recorded scene, the increase of the *Z* coordinate (scene depth) of observed points is small relative to its value. Then the projection of movement of any point of the surrounding scene on the camera sensor plane is a two-dimensional vector $[u \ v]^T$, where [3]:

$$\begin{cases} u = x_K - x_P = \dfrac{fX_K}{Z_K} - \dfrac{fX_P}{Z_P} \approx f\left(\dfrac{\Delta X}{Z_P} - \dfrac{X_P \cdot \Delta Z}{Z_P^2} \right) \\[4mm] v = y_K - y_P = \dfrac{fY_K}{Z_K} - \dfrac{fY_P}{Z_P} \approx f\left(\dfrac{\Delta Y}{Z_P} - \dfrac{Y_P \cdot \Delta Z}{Z_P^2} \right) \end{cases} . \tag{6}$$

After substituting the increments of coordinates from (4) to (6) one obtains [4]:

$$\begin{cases} u = \dfrac{-Uf + xW}{Z} + \alpha \dfrac{xy}{f} - \beta\left(\dfrac{x^2}{f} + f \right) + \gamma y \\[4mm] v = \dfrac{-Vf + yW}{Z} + \alpha\left(\dfrac{y^2}{f} + f \right) - \beta \dfrac{xy}{f} - \gamma x \end{cases} . \tag{7}$$

The *P* index has been omitted because the initial position is always associated with the first of two consecutive frames of an analyzed sequence.
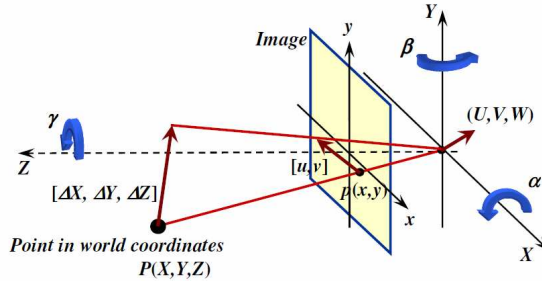

Fig. 1. Projection of the movement of the point P in the 3D scene on the camera sensor plane.

In the canonical stereoscopic system the measure of the scene depth (*Z* coordinate) is calculated from the disparity, i.e. the displacement of the image of a point in the left camera relative to its image in the right one. The dependence of *Z* on the disparity *d* is given by:

$$Z = \frac{fB}{d}, \tag{8}$$

where *B* (basis) is the distance between the optical axes of the cameras.

After substituting (8) to (7), the image motion vector of any point in the static scene is described by the following dependence:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} -\dfrac{d}{B} & 0 & \dfrac{xd}{fB} \\[3mm] 0 & -\dfrac{d}{B} & \dfrac{yd}{fB} \end{bmatrix} \begin{bmatrix} U \\ V \\ W \end{bmatrix} + \begin{bmatrix} \dfrac{xy}{f} & -\left(f + \dfrac{x^2}{f} \right) & y \\[3mm] \left(f + \dfrac{x^2}{f} \right) & -\dfrac{xy}{f} & -x \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}, \tag{9}$$

which is the sum of the translational and rotational component of video system movement in the scene. This allows an initial estimation of the motion vectors (motion field) of characteristic points in the image on the basis of the angular velocity and linear acceleration

measured by inertial sensors rigidly connected to the camera system. Predicting this movement is important in the case of the intended application in a head-mounted stereovision system, where the main causes of motion in the image are the rotations of the user's head. Estimation of the camera egomotion vector requires a number of image analysis operations:
- colour image conversion to greyscale, distortion correction and rectification,
- detection of characteristic points – image corners and determination of their motion [3],
- estimation of the disparity between the position of these points in the left and right image of stereoscopic pair with the use of a block matching technique [5],
- estimation of corner displacements between consecutive images in the sequence.

Image characteristic points, in this case corners, are detected with a standard Harris detector. More computationally efficient detectors were developed in recent years, such as SIFT or SURF [6]. They could replace the Harris detector, but the time reduction of the total algorithm would not be significant due to the other, more time consuming steps. Searching for every corner displacement between successive frames in the sequence $[u\ v]^T$ is realized by matching image blocks in a two-dimensional neighbourhood. The full search algorithm in a given neighbourhood of the point was implemented. The block matching procedure in two dimensions is the most time-consuming part of the whole algorithm and greatly benefits from the limitation of the search range. Initial estimation of the motion field in the image based on the readings of electronic, inertial sensors helps to predict the region of the search without increasing the computational effort.

3. Algorithm for estimation of the egomotion vector of the camera system

Estimation of the camera egomotion vector is performed using a modified random sample consensus (RANSAC) algorithm. This is an iterative algorithm that allows to find the model parameters describing the measurements in the presence of noisy data or data not meeting the assumptions of the model (outliers). In the application of identification of the camera movement model parameters the algorithm reduces errors resulting from erroneous disparity or motion field estimation in the image, and due to the presence of moving objects in the scene [7]. The implemented algorithm operates according to the following scheme:

```
Repeat N times:
{
        select random  data from a set;
         repeat M times:
        {
                determine the model based on the selected data;
                for the next step select the data matching the model with an error less then T;
                if the number of data is less than the threshold D, repeat the draw,
                if the model is better than the one found previously, remember it;
        }
}
```

In the application of camera egomotion estimation, measurement data are composed of the sets of motion vectors and the disparity of characteristic points in the image $[x\ y\ u\ v\ d]^T$. The model is represented by a vector of egomotion parameters $[U\ V\ W\ \alpha\ \beta\ \gamma]^T$. The parameters of the algorithm, arrived at after a series of test sequences, are set as follows: N=5, M=5, T=1[pixel], D=30. Estimation of the model parameters is done by minimizing the mean square error between the estimated and the measured motion of characteristic points in the image.

4. Estimation of the gravity vector

To allow the initial estimate of motion in the image $[u\ v]^T$ on the basis of signals from the inertial sensors it is necessary to estimate the gravity vector **g** in the reference frame associated with the cameras. After substitution of this vector from the triaxial accelerometer measurements, the $\boldsymbol{a}_{INS}$ vector of dynamic linear acceleration **a** of the acquisition system is obtained, which allows to determine its speed [8]. An additional benefit of separating the vector **g** is the ability to classify one of the planes found in the disparity image as the ground plane. This information can be used to alert the blind user of the obstacle avoidance system of any faults or changes to this plane (e.g., curbs, stairs or holes). The structure of the gravity estimation algorithm (Fig. 2) follows known solutions of inertial navigation [1, 7]. The current estimate of vector **g** is rotated by the angles measured by the triaxial gyro. To avoid the accumulation of angular velocity measurement errors, in successive runs of the algorithm (in each frame of image sequence) the current estimate of the vector **g** is modified by adding the weighted value of the current measurement of $\boldsymbol{a}_{INS}$.
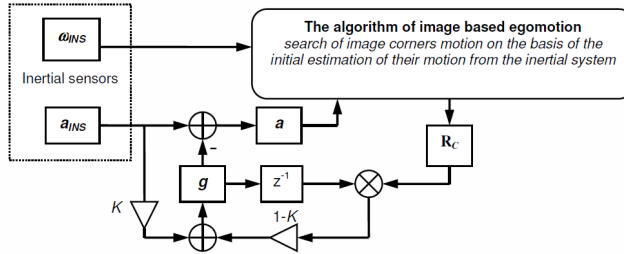

Fig. 2. Diagram of the algorithm for the gravity vector extraction.

Weighting is done through multiplication by a constant $K \ll 1$. The more accurate the estimate of the average angular velocity between successive images of the sequence, the smaller may K be. In experiments performed on synthetic data K could be set to 0.001. In the real acquisition system K values in the range <0.01, 0.02> gave the best results.

5. Integration of image data and signals from inertial sensors in egomotion estimation

The starting point of the presented work was the algorithm for egomotion estimation only on the basis of visual data [5]. The inclusion of an inertial sensor system reduced the time-consuming computations of the image motion field prediction. It also allowed a better estimation of camera motion in the case of low image quality due to poor lighting or as a result of image blurring from rapid camera rotations.

Initial estimation of motion in the image consisted of calculating the coarse motion vector $[u\ v]^T$ according to (9), based on the measured angular velocity. In the simulation experiment an inclusion of linear speed also gave good results. It was modified according to the estimated acceleration **a** in each image frame. In the experiment with real acceleration sensors estimation of **a** was not sufficiently precise and the prediction was calculated only on the basis of signals from the gyroscopes. Effectively, it was possible to narrow the search range of corner displacements by imaging methods to 10 points around the predicted positions, although the actual motion in the image often exceeded 30 pixels. This approach has allowed to reduce the computational cost of motion field search about ten times without noticeable deterioration in the algorithm's effectiveness.

Further improvement of the egomotion estimation accuracy was achieved by weighted summation of the motion vectors obtained from video and inertial sensor based algorithms:

$$
\begin{bmatrix} U \\ V \\ W \\ \alpha \\ \beta \\ \gamma \end{bmatrix} = cert \cdot \begin{bmatrix} U \\ V \\ W \\ \alpha \\ \beta \\ \gamma \end{bmatrix}_{VIS} + (1 - cert) \cdot \begin{bmatrix} U \\ V \\ W \\ \alpha \\ \beta \\ \gamma \end{bmatrix}_{INS} , \tag{10}
$$

where: *VIS* – index of the motion vector estimated from image data, *INS* – index of the motion vector estimated from measurements of linear acceleration and angular velocity, *cert* – measure of certainty of the motion vector estimation from the image data.

The weights of both sources were adjusted dynamically on the basis of measurement certainty. The error rate of the inertial system does not depend on the position, orientation, or other factors associated with the movement. Camera egomotion estimation from image data is sensitive to many things: the camera motion, the movement of objects in the observed scene, content of the images, the lighting conditions and even the possible masking of one of the cameras. In these situations the number of points moving in the images according to the estimated motion model is reduced. The ratio of the number of corners meeting the model to the total number of corners found was used as the measure of certainty *cert* . Weighted summing defined in (10) improved the egomotion estimation accuracy in situations of incorrect image acquisition.

Obviously, more advanced techniques of data fusion from various sources can be found. Some of the frequently used approaches use Kalman filtering [9, 10] and the authors plan to implement also this technique for comparison. Other methods, such as stochastic filtering [11] or particle filtering [12] are very time consuming, so the main advantage of the solution developed by the authors is the significantly lower computational complexity.

6. Experimental verification of the developed algorithm

6.1. Algorithm validation on synthetic data

In order to verify the correctness and evaluate the effectiveness of the developed algorithm, synthetic image sequences with simulated inertial sensor signals have been prepared. The image sequences were created using the POV-Ray 3D modeling environment. An urban scene was prepared and filled with typical objects that would be encountered by a pedestrian user of the system (Fig. 3).

a)   b)

Fig. 3. A single frame of image sequence of a synthetic scene a) with the camera aligned with the axes of the external reference frame, and b) with the camera after its translation and rotation.

The trajectory of the camera was defined and featured simultaneous forward motion and rotation. The following parameters of the simulated stereo system were assumed: canonical stereo system with pinhole cameras, base $B = 0.1$ m, focal length $f = 443$ pixels, and 512x384 image size. A sequence of one hundred images with a frame rate equal to 10fps was generated. Artificial noise was added to the simulated inertial sensors at a level comparable to that of the real sensors, integrated in a constructed module and described in [13]. At the same time a file was created that contained information about the right camera position relative to the static frame of reference for each frame of the sequence. This allowed the determination of the displacement estimation error as a function of the frame number.

An application for quantitative evaluation of the effectiveness of the developed algorithm was developed in the MATLAB environment. It allowed to display the measurements from the accelerometers, the estimated dynamic acceleration, the gravity vector (Fig. 4), the estimated motion parameters (Fig. 5), the movement trajectory and localization errors (Fig. 6). Ideal sensor readings (without interference) were generated as a reference, in order to validate the simulated measurements and the accuracy of their reading in the application running the motion estimation algorithm. The algorithm was set to estimate egomotion on the basis of the inertial system only. The K factor was set to a very low value ($K = 0.0001$). The result of motion estimation in these conditions is shown in Fig. 6. Despite the complex nature of the analyzed motion, the localization error was no greater than 1% of the traveled distance.
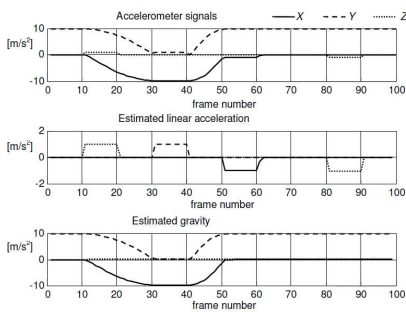

Fig. 4. The result of the algorithm for separation of the gravity vector and the dynamic accelerations of acquisition system as a function of frame number.
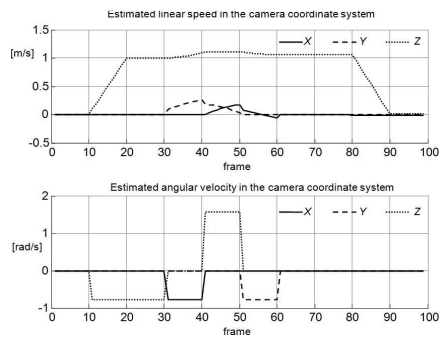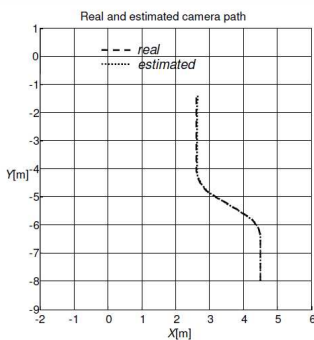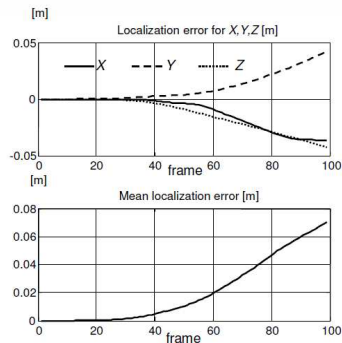

Fig. 5. Plots of estimated motion parameters as a function of frame number.

(a)

(b)

Fig. 6. Comparison of the real and estimated path of the stereovision module a), and localization errors b) for purely inertial navigation using ideal sensors.

The introduction of noise at a level comparable to what can be expected in real sensors produces a strong increase in the localization error and its accumulation in time. The introduced noise possessed a normal distribution with a zero mean value and a standard deviation of 0.2% of the full measurement range. For a sequence of 100 images the localization error often exceeded 50% of the traveled route, which eliminates the possibility of using the inertial sensors alone for egomotion estimation in the obstacle avoidance system.

The vision based algorithm for egomotion estimation was also inaccurate in the same conditions (Fig. 7). The algorithm suffered lack of robustness during rapid rotations of the cameras, which in turn led to an incorrect estimation of the direction of motion and accumulation of the localization error. The reason for such large errors was the limit of the image motion field search to 10 pixels and the existence of similar regions, falsely matched in the image motion search procedure. Increasing the motion field search range is not possible, because it greatly increases the demand for computing power, which is squarely proportional to the search range. False fits of image regions occur for scenes containing repetitive patterns and their likelihood increases with increased search range.

A fusion of inertial sensor signals and image data proved to be an attractive alternative to the vision based egomotion algorithm. After refinement of the parameters of the implementation of the data fusion algorithm, a significant improvement in determination of the camera path was obtained (Fig. 8). Using the developed data fusion algorithm for egomotion estimation decreased the localization error six times in comparison to the error of the purely vision-based egomotion estimation.
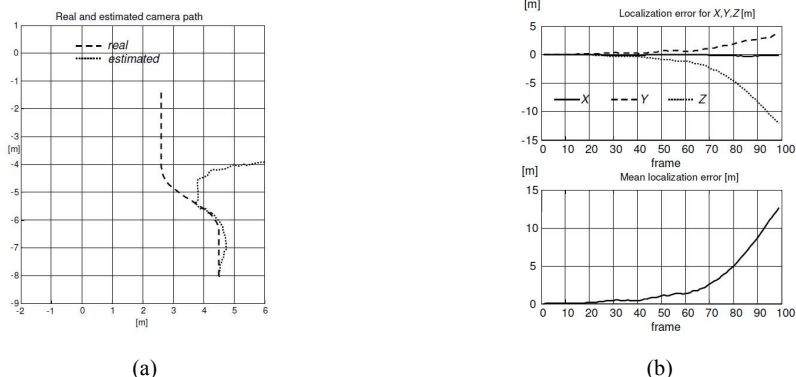

(a) (b)

Fig. 7. Comparison of real and estimated path of the stereovision module a) and localization errors b) for a purely vision-based egomotion algorithm.
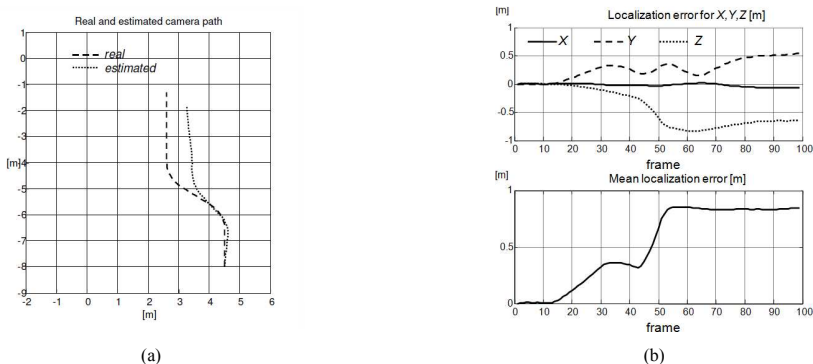

(a) (b)

Fig. 8. Comparison of real and estimated path of the stereovision module a) and localization errors b) for the data fusion-based egomotion estimation algorithm.

6.2. Algorithm verification on real data

The precision of the data fusion-based algorithm was also measured for real scenes recorded by the constructed stereovision system [13]. The following parameters were used: B=0.086 m, camera focal length f = 435 pix, size of images 512x384 pix, frame rate 12 fps. A stereovision headmount with digital Pointgrey FLEA cameras was used. A digital video stream, produced by the cameras, was captured to a standard PC.

To measure the acceleration and angular velocity the following inertial sensors were used: two integrated, two-axis MG1101B gyroscopes produced by GYRATION and a three-axis linear acceleration sensor 3LS02AL from STMicroelectronics. Registration of signals from the sensors was synchronized with the acquisition of images. The measurement range of the linear acceleration sensor is +/-2 g for each axis. The electronic gyroscopes measure the angular velocity in the range of +/-180deg/s.

The camera was moving on a curved line and was subjected to rotations. The actual motion track has been measured by photogrammetric methods [14], by analyzing the recorded images of a model scene composed of regular boxes (Fig. 9a).
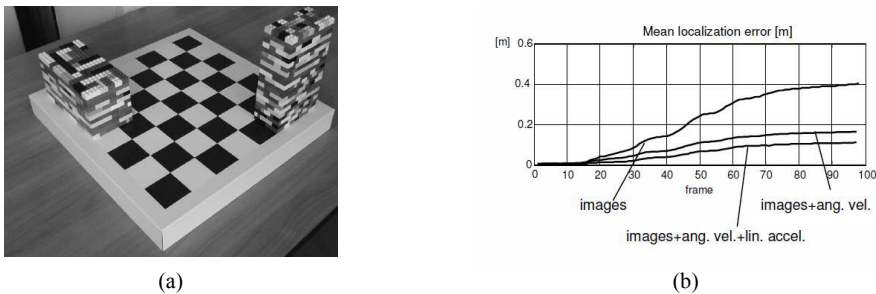

| (a) | (b) |

Fig. 9. Test scene a) and comparison of localization errors for only visual information, and after fusion of the image and sensor data b).

Signals from the inertial sensors of the real system demonstrated good performance of the data fusion-based algorithm for egomotion estimation. The tracking accuracy has been improved 4 to 5 times over the algorithm using purely visual methods (Fig. 9b). In addition, the estimation of the gravity vector **g** was accurate enough to recognize the horizontal plane. The estimation of the separated dynamic acceleration was not very precise due to measurement errors and dynamic properies of the accelerometer, so the calculated speed could not be used in (9) to predict the image motion field. Still, it was used in the data fusion formula (10) and provided an improvement of the egomotion estimation (Fig. 9b).

7. Conclusions

The integration of image information, angular velocity and linear acceleration allowed to efficiently track the position changes of the navigation system in a static scene. In the proposed system, the main cause of the observed motion in the images is the rotational component of the camera motion in the scene. The use of gyroscopes gave the opportunity for rough estimation of the motion of points in the image, allowing to reduce the range of the motion field search. The use of accelerometers enabled the estimation of the gravity vector, which will be used to recognize the ground plane, which is important from the viewpoint of supporting a blind pedestrian in independent travel [15, 16].

Acknowledgements

References

[1] Titterton, D.H, Weston J.L. (2004). *Strapdown Inertial Navigation Technology*. Progress in Astronautics and Aeronautics, MIT Lincoln Laboratory, 2nd. Edition.

[2] Ma, Y., Soatto, S., Kosecka, J., Sastry, S. (2003). *An Invitation to 3D Vision*. Springer.

[3] Bruss, A.R., Horn, K.P. (1983). Passive Navigation. *Computer Vision, Graphics, and Image Processing 21*, 3-20.

[4] Argyrost, A.A., Orphanoudakis, S.C. (1997). Independent 3D Motion Detection Based on Depth Elimination in Normal Flow Fields. *Comp. Vision and Pat. Recogn. Conf. (CVPR'97)*, San Juan, Puerto Rico.

[5] Skulimowski, P., Strumillo, P. (2008). Refinement of depth from stereo camera ego-motion parameters. *Electronics Letters,* 44(12), 729-730.

[6] Bauer J., Sünderhauf N., Protzel P. (2007). Comparing Several Implementations of Two Recently Published Feature Detector. *In Proceedings of the International Conference on Intelligent and Autonomous Systems, IAV, Toulouse, France, 2007.*

[7] Dante, A., Brookes, M. (2003). Precise real-time outlier removal from motion vector fields for 3D reconstruction. *International Conference on Image Processing,* I- 393-6, 1.

[8] Foxlin, E.M., Harrington, M., Altshuler, Y. (1998). Miniature six-DOF inertial system for tracking HMDs. *Proc. SPIE 3362*, 214-228.

[9] Veth, M., Raquet, J. (2007). Fusion of Low-Cost Imaging and Inertial Sensors for Navigation. *NAVIGATION - Journal of the Institute of Navigation,* 54(1).

[10] Armesto, L., Tornero, J., Vincze, M. (2007). Fast Ego-Motion Estimation with Multi-rate Fusion of Inertial and Vision. *The International Journal of Robotics Research*, 26(6), 577-589.

[11] Jones, E.S., Soatto, S. (2011). Visual-Inertial Navigation, Mapping and Localization: A Scalable Real-Time Causal Approach. *International Journal of Robotics Research*, 30(4), 407-430.

[12] Baranski, P., Polanczyk, M., Strumiłło, P. (2011). Fusion of Data from Inertial Sensors, Raster Maps and GPS for Estimation of Pedestrian Geographic Location in Urban Terrain. *Metrology & Measurement Systems,* 18(1), 145-158.

[13] Pełczyński, P., Ostrowski, B., Rzeszotarski, D. (2009). Mobile System for Passive Navigation in 3D Scene. VIII *KKE 2009, East Darłówko,* 682-687. (in Polish)

[14] Pełczyński, P. (2011). Model Based Estimation of Camera Position in 3D Scene. The *3rd International Conference on Image Processing & Communications, Bydgodszcz.*

[15] Skulimowski, P., Strumiłło, P. (2007). Obstacle Localization in 3D Scenes from Stereoscopic Sequences. *15th European Signal Processing Conference (EUSIPCO 2007), Poznan, Poland,* 2095-2099.

[16] Bujacz, M., Skulimowski, P., Strumiłło, P. (2011). Sonification of 3D scenes using personalized spatial audio to aid visually impaired persons. *International Conference on Auditory Display, Budapest, Hungary, 2011.*