

Multi-focal Feature Tracking for a Human-Assisted Mobile Robot

Tingting Xu¹, Yi Guo¹, Kolja Kühnlenz^{1,2} and Martin Buss¹

¹Institute of Automatic Control Engineering (LSR)

²Institute for Advanced Study (IAS)

Technische Universität München

D-80290 Munich, Germany

Email: {tingting.xu, kolja.kuehnlenz, m.buss}@ieee.org, yi.guo@mytum.de

Abstract— In the project *Autonomous City Explorer*, an interactive robot is designed to find its way to a given destination in unknown urban environments by interacting with pedestrians. Considering applications in a human dominated environment, the robot can be sent to a destination by tracking a landmark selected by users and described by 2D image features. To achieve a natural landmark selection from the user perspective and an accurate feature tracking for a safe robot navigation, the robot preselects visual features and presents the users only the image regions providing higher tracking accuracy. Furthermore, a multi-focal camera system is used to extend the sensing range. SIFT, Harris corner and optical flow used for tracking and self-localization are compared and applied to different visual sensors. A coordination strategy is realized, in which the camera with wide field of view is used for robot orientation control and the high-resolution camera is applied for robot forward motion control. The performance is experimentally evaluated.

I. INTRODUCTION

In recent years, interactive robots have become a focus of robotics research. Robots have been developed to assist humans in household activities [1] [2], or guide humans through museums [3] [4] [5]. Considering applications in a human dominated environment, a typical task for robot operation is fetch and carry. For instance, a user sends the robot to a certain destination. Then, the robot searches in that area for target objects [6] and performs manipulation tasks.

In the *Autonomous City Explorer (ACE)* project (see Fig. 1), we created a robot that is able to navigate in unknown urban environments autonomously without use of GPS or map knowledge and find its way to a given destination using information from pedestrians such as direction and/or distance [7].

Compared to other interaction modalities, visual information can describe a destination more conveniently than gesture regarding user performance, more simply than verbal information, more flexibly than map knowledge, and more naturally than metric input. Moreover, we apply a multi-focal vision system to extend the sensing range and strengthen the navigation performance. Multi-focal vision systems have already been applied in many robotics domains such as surveillance systems [8], visual attention systems [9] [10] [11] [12] [13] or visual servoing systems [14] etc.

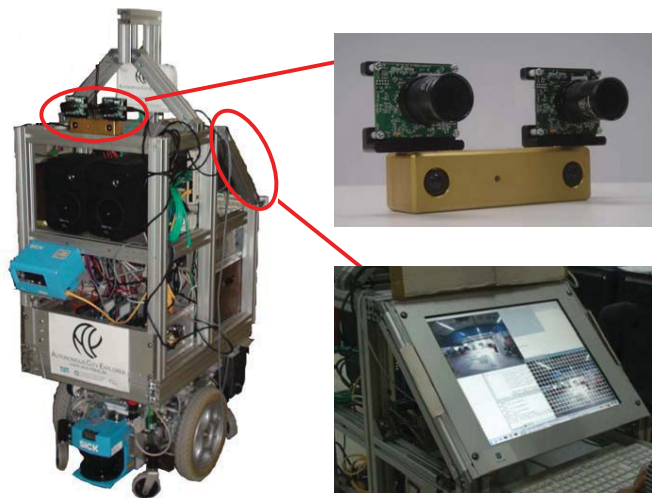


Fig. 1. The ACE robot (left) with a multi-focal vision system (right top) and a touch screen (right bottom).

In most applications, low-resolution visual sensors provide an overview about the environment. The image region containing target objects or features of interest is focused by the high-resolution visual sensors. However, no simultaneous control of robot motion using data from both sensors is applied. Moreover, we consider a mobile robot applied in a dynamic environment. The destination can disappear, if it is hidden by humans occasionally or the robot changes its orientation to avoid obstacles. In addition, user convenience is also an issue we should consider.

To deal with the problems such as possible loss of the tracked information caused by dynamic environment and a safe navigation over a relatively long distance, a multi-focal vision system containing a wide-angle stereo camera and two telephoto cameras is used. Given an input image, a user selects a landmark in the destination area that the robot is supposed to approach. The landmark is selected in a wide-angle image in order to keep the desired destination in the input image. This 2D landmark should be tracked continuously during the robot movement, while the relative position between the landmark and the robot itself in 3D space should also be updated continuously. Therefore, land-

mark re-detection in real time based on low-resolution visual data is also essential in addition to highly accurate tracking using high-resolution visual data. SIFT, Harris corner and optical flow technique are applied to different sensors. A coordinated control of robot motion is realized in which the wide-angle camera is used for robot orientation control and the telephoto cameras are responsible for robot forward motion control.

The remainder of this paper is organized as follows: In Section II an overview of system hardware is given. In Section III landmark tracking and position estimation using the multi-focal camera system are described in detail. Experimental results are presented and discussed in Section IV. In Section V conclusions are given.

II. HARDWARE DESCRIPTION

A. The Mobile Robot and its Components

The mobile robot *ACE* comprises a differential drive mobile platform by BlueBotics SA, two laser range finders for obstacle avoidance, a loud speaker, a touch screen and a multi-focal camera system (see Fig. 1). The maximum velocity is 1.4 m/s, the maximum acceleration 1.35 m/s². Two PCs are placed on the top of the platform, one for navigation/interaction and the other one for image processing. Further information can be found in www.ace-robot.de.

B. Multi-focal Vision System

The navigation destination is given by users via landmark selection in a 2D image. As mentioned in Section I, to ensure the feature selection and tracking success as well as an accurate position estimation of the destination, a multi-focal camera system is needed.

A preliminary multi-focal camera configuration is decided (see Fig. 1 right top). In order to avoid extrinsic parameter calibration error propagation, two dragonfly telephoto cameras with focal lengths of 12 mm each are rigidly mounted on a Bumblebee stereo camera of Point Grey Research Inc. with focal lengths of 2 mm each currently. In this paper, only the left wide-angle camera is used.

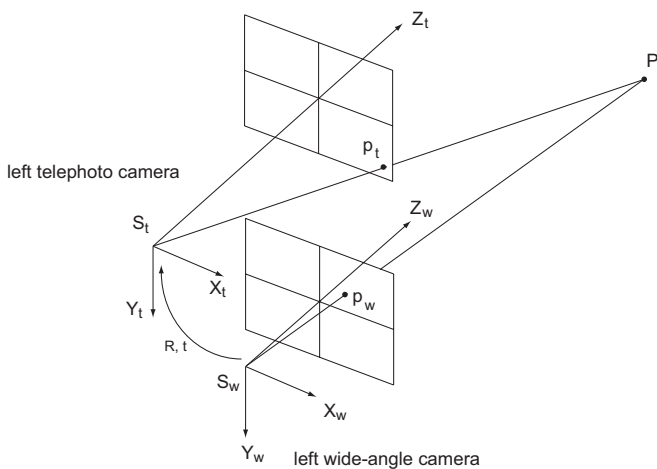


Fig. 2. The multi-focal camera model and the frame definition.

Fig. 2 illustrates the multi-focal camera model and the camera frame definition. The left wide-angle camera is regarded as the reference. The left telephoto camera frame S_t and the left wide-angle camera frame S_w are coincident, since the optical axes of the cameras are assumed to be parallel. Therefore, the rotational matrix \mathbf{R} is a unit matrix, while the translational vector $\mathbf{t} = (t_x, t_y, t_z)^T$ indicates the displacement of the left telephoto camera frame with respect to the left wide-angle camera frame. A 3D point P in front of the vision system has the projection p_t on the telephoto image and the projection p_w on the wide-angle image. Since a direct feature matching between images with different resolutions is not sufficiently reliable referring to the scale space, which has been experimentally tested using SIFT features, the projection of landmark selected in wide-angle image data on telephoto images should be achieved using cameras geometric relationship. Given a 2D point $p_w = (x_w, y_w)$ in wide-angle image, selected at first by a user, the corresponding 2D point $p_t = (x_t, y_t)$ on the telephoto image is to be determined.

As known, the general form of the *camera calibration matrix* \mathbf{K} is

$$\mathbf{K} = \begin{pmatrix} \alpha & \gamma & x_0 \\ 0 & \beta & y_0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (1)$$

where α and β represent the focal lengths of a camera in terms of pixel dimensions in horizontal and vertical directions in 2D image-space, respectively, γ the skew parameter, and (x_0, y_0) the principle point equal zero.

Then the projections of the 3D point P with coordinate $(X, Y, Z)^T$ onto the wide-angle and the telephoto images are represented by:

$$\lambda \begin{pmatrix} x_w \\ y_w \\ 1 \end{pmatrix} = \begin{bmatrix} \alpha_w & \gamma_w & 0 & 0 \\ 0 & \beta_w & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}, \quad (2)$$

$$\lambda \begin{pmatrix} x_t \\ y_t \\ 1 \end{pmatrix} = \begin{bmatrix} \alpha_t & \gamma_t & 0 & 0 \\ 0 & \beta_t & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{R} & -\mathbf{R}\mathbf{t} \\ 0 & 1 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}, \quad (3)$$

where λ is a scale factor and has the value Z .

We assume $\gamma_t = \gamma_w = 0$, since they are normally much smaller than α_t , α_w , β_t and β_w . Therefore,

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \frac{1}{Z - t_z} \left[\begin{pmatrix} Z\alpha_t/\alpha_w \cdot x_w \\ Z\beta_t/\beta_w \cdot y_w \end{pmatrix} - \begin{pmatrix} \alpha_t t_x \\ \beta_t t_y \end{pmatrix} \right]. \quad (4)$$

Since

$$t_x \ll Z \quad \text{and} \quad t_z \ll Z, \quad (5)$$

we simplify the Eq. 4 into

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} \alpha_t/\alpha_w \cdot x_w \\ \beta_t/\beta_w \cdot y_w - \beta_t t_y/Z \end{pmatrix} \quad (6)$$

by assuming $t_x = 0$ and $t_z = 0$.

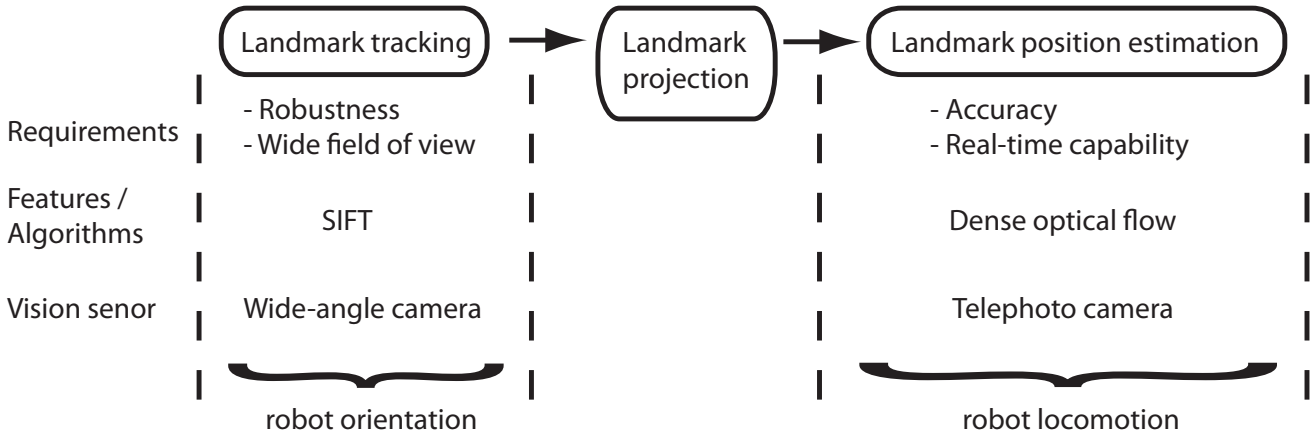


Fig. 3. Feature tracking procedure with specified requirements, appropriate features/algorithms and suitable visual sensors.

What should be pointed out is that all the analysis above does not consider the effects of lens distortion. For telephoto cameras, which are equipped with lenses of large focal length, the distortion effect is not noticeable and can be ignored. For the wide-angle camera, significant distortion is exhibited, especially at the image borders. Besides an accurate calibration for the multi-focal camera system, the robot orientation is controlled using the wide-angle image data to locate the selected landmark into the image center, in order to reduce the influence of the distortion.

III. MULTI-FOCAL FEATURE TRACKING

Fig. 3 illustrates the multi-focal feature tracking procedure. The selected landmarks are represented by 2D features. In the landmark tracking step, features are supposed to be tracked robustly during the robot motion. In the landmark position estimation phase, the position of the tracked features is updated until the robot arrives the desired position. As mentioned in Section I, one of the requirements for a robust landmark tracking is a wide view field of the vision sensor, while the accuracy has a high priority in the position estimation. Therefore, the wide-angle camera and the telephoto cameras are used for landmark tracking and position estimation respectively. The landmark projection from wide-angle images onto telephoto images is accomplished as described in Section II.

For each step the appropriate 2D feature is to be determined according to the requirements. Three conventional features/algorithms are considered, namely SIFT, Harris corner and dense optical flow technique.

A. Landmark Selection

To guarantee the tracking success, the selected image region should have distinctive and robust features in the 2D input image. Only image regions containing features are supposed to be selected by the user. Therefore, 2D image features should be calculated before the user selects the landmark on the input image. Thereby, the input image is divided into blocks consisting of 32×32 pixels each. Then, the feature number in each block is computed. Only

the image regions with relatively high feature number are provided as candidates to the user. If the selected image region does not contain enough features, complementary modules such as a metric input may be needed, which is one of the future works. But usually, humans describe a destination using the phrase like “look for a book on that bookshelf over there” or “go to the square in front of that church”. Empirically, most outdoor destinations contain many features.

After the image region selection, the robot rotates itself to locate the landmark in the wide-angle image center. Considering the assumptions and approximation made in Section II, it is more accurate for the projection onto telephoto images if the selected landmark is located near the image center of the wide-angle images. Therefore, during the following tracking process, the wide-angle image data is used to control the robot orientation, in order to keep the selected landmark in the image center.

B. Landmark Tracking Using Wide-Angle Camera

During robot locomotion, the features of the landmark are the focus of robot attention and should be tracked robustly. Preconditions for a successful tracking are as follows:

- Wide field of view is needed to keep the destination in the view consistently;
- Features should be robust against scale change, view point change, and illumination change;
- Features should be re-detectable, for the case if they are lost in the field of view due to other objects covering the cameras or obstacle avoidance.

Therefore, we use the wide-angle camera to track the landmark and solve the problem how to represent the selected destination to achieve a robust tracking behavior. Due to the reasons mentioned before, Scale Invariant Feature Transformation (SIFT) features [16] are used here to describe the landmark robustly. Fig. 4 left shows an example of SIFT feature matching result. The solid lines are the matching results between two input images along robot locomotion. To delete wrong matches, namely the highlighted dashed corresponding lines in the left image, the RANdom SAMple

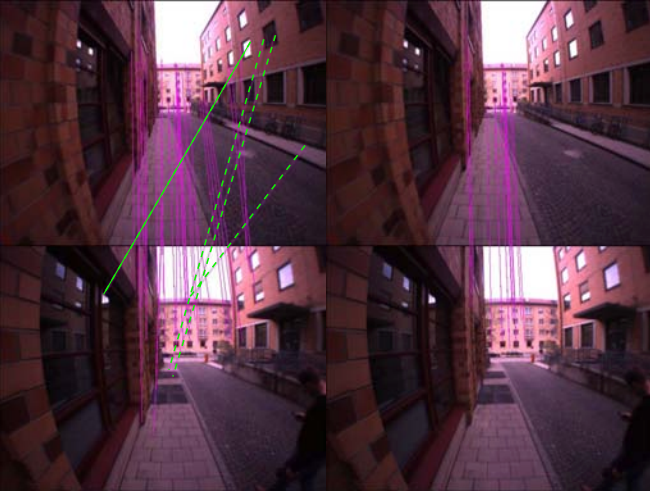


Fig. 4. SIFT feature matching result of two consecutive input images before (left) and after (right) using RANSAC Algorithm

Consensus (RANSAC) algorithm [17] is applied after feature matching (see Fig. 4 right).

As known, the more features are extracted, the more robustly can the landmark be represented. However, to extract and match many features also means a high time cost. To lower the computation cost, we only extract the features in a *feature window* around the selected image region (see Fig. 5). A comparison of tracking performance using *feature windows* with different sizes is conducted in Fig. 6. Using a window of 96×96 pixels, number of matched features between two consecutive images is large, while the time cost is also high. Using window size 48×48 , the computation in average is not expensive. However, loss of features happens. For our experiments, the *feature window* is chosen to be 64×64 pixels, which is a compromise between robustness and real-time performance.

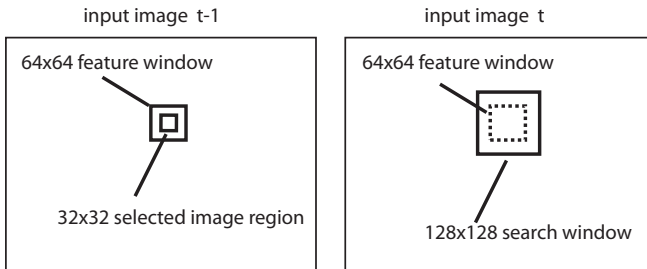


Fig. 5. 32×32 selected image region, 64×64 feature window and 128×128 search window defined on wide-angle input images.

To lower the computational time for feature matching, we also predict a *search window* of 128×128 pixels for the consecutive image according to estimated current robot motion. If the destination is not located in this predicted window, which means either the destination is occluded or the robot motion has changed a lot, the *search window* is doubled, until it reaches the original size of the input image. In a word, the SIFT features extracted in the previous *feature window* are matched with the features extracted in

the current *search window*. Thereby, the robot is always controlled toward the center point of the features. The center of gravity of the matched SIFT features in the search window at the time step t is saved. At the time step $t + 1$, the features in the feature window around this center are used to be matched with features in the search window of image $t + 1$. In this way, the SIFT features are updated at each time step, such that the feature number does not get smaller and smaller, in order to make a feature matching over a large workspace possible.

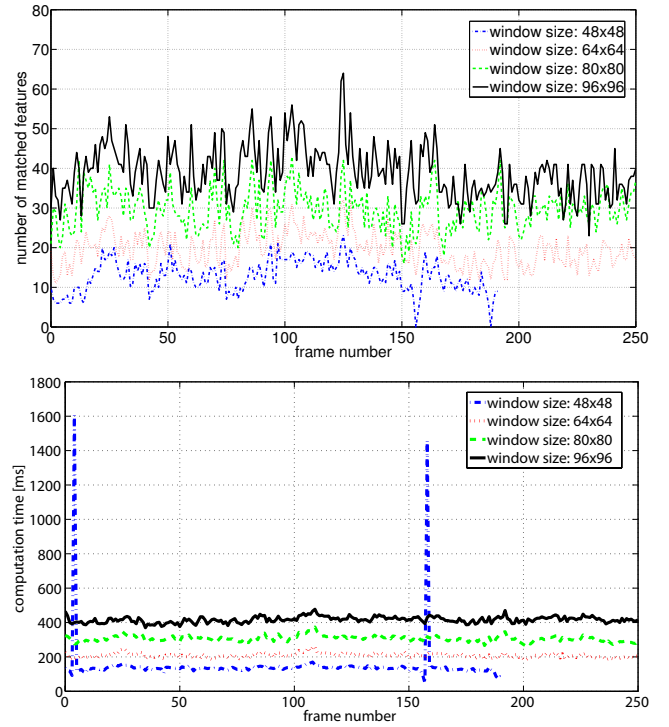


Fig. 6. The feature tracking performance using feature windows of different sizes. Top: number of matched SIFT features in the window; Bottom: computational time per frame.

C. Landmark Position Computation using Telephoto Cameras

Afterwards, mapping the image region containing features found in the wide-angle view onto telephoto images is conducted. This image region in the wide-angle images is enlarged and further processed in the telephoto images.

In order to achieve an accurate position estimation we use the high-resolution telephoto images to compute the relative position between the landmark and the robot. For stereo cameras the stereo triangulation is normally used to reconstruct a 3D point from its perspective projection on the image planes of the cameras, if the relative position and orientation of the two cameras are available. There are various algorithms to find the corresponding pixels of the projected image region in the left and right telephoto camera images such as SIFT feature matching, Harris corner matching or optical flow computed using Lucas-Kanade algorithm [18] [19] etc. SIFT feature matching is more robust than the other

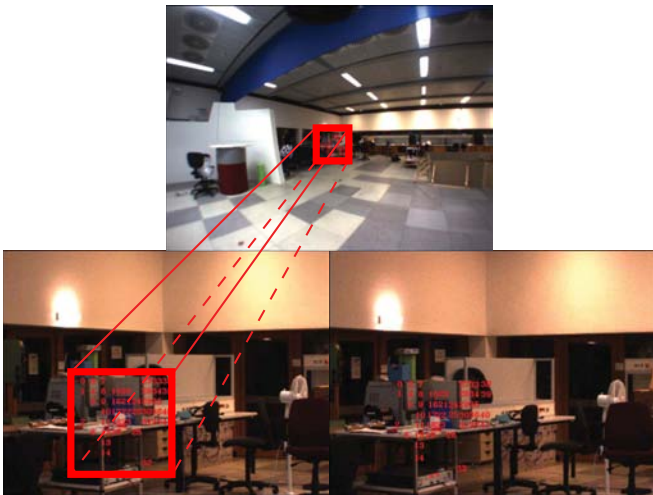


Fig. 7. Projection of image regions selected on the wide-angle images onto the telephoto images.

two algorithms. However, compared to the 64×64 feature window used in the wide-angle images (see Fig. 7 upper image), its projected image region on the telephoto input image is much larger (see Fig. 7 lower image). Therefore, SIFT is not suitable at this step in terms of real-time ability. Although the computation cost is low, Harris corner detection is not robust enough. In two consecutive image regions of 120×120 pixels, only about 20 Harris corner features are detected and matched (see Tab. I). The optical flow technique using Lucas-Kanade algorithm is chosen due to relatively robust feature matching and low computational cost.

Technique	number of matched point	computation time
RANSAC-SIFT	50	500 ms
Harris corner	20	42 ms
Optical flow	40	40 ms

TABLE I

Comparison of matched point number and computational cost using three different features/algorithms in two consecutive 120×120 search windows.

We assume that the landmark lies on a planar surface, since the relative distance between any two points on the landmark in the robot moving direction is much smaller than the relative distance between a 3D point on the landmark and the robot. The orientation of the robot is estimated using wide-angle visual data by integration of the relative orientations at each former time step, while the forward movement and position of the robot are estimated using telephoto image data. Relying on the telephoto image data, the robot is controlled to move forward until it reaches the destination.

IV. EXPERIMENTS AND DISCUSSIONS

The performance of our vision system is evaluated experimentally. The robot shown in Fig. 1 was placed in a

room of approximately $15m \times 15m$ and should stop 3 m in front of the selected landmark. Experiments with different destinations and different distances were conducted.

Fig. 7 illustrates the image region selected by a user (top) on the wide-angle image and its projection on the left telephoto image (left bottom). The corresponding points between the left telephoto image and the right telephoto image (right bottom) are also shown. The numbers shown in the telephoto images are the pixels matched using the Lucas-Kanade method.

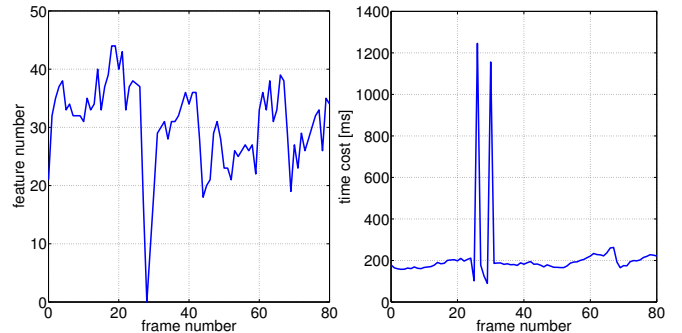


Fig. 8. SIFT feature tracking and re-detection performance during robot locomotion. Left: the number of matched SIFT features per frame; Right: the computational time per frame

Fig. 8 shows the experimental result of SIFT feature tracking using the wide-angle camera. Approximately 30 features were matched between the *feature window* in the previous image and the predicted *search window* in the current image. About 200 ms is used for SIFT feature extraction and matching. At 28-th frame the landmark was hidden by a human and not seen by the robot. A feature re-detection was conducted in the whole input image successfully which caused a six times longer computation time.

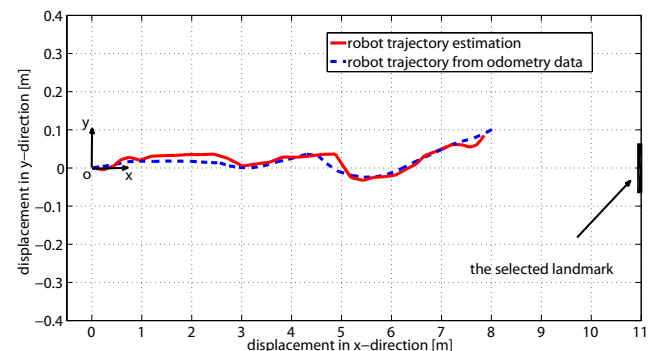


Fig. 9. Bird's eye view of the robot's path. The solid line indicates the results of estimation using the multi-focal vision system. The dashed line indicates the robot movement estimated using wheel encoders.

The relative position estimation was conducted satisfyingly, as illustrated in Fig. 9. The robot was approximately 11 m far from the landmark at the beginning and moved over a distance of 8 m. The solid line is the result estimated using vision data, while the dashed one is the result estimated by wheel encoders. Differently from visual SLAM, the position of the selected landmark in our experiments was not known

at the beginning. To obtain ground truth, we measured the relative position between the landmark and the initial position as well as the final position of the robot.

However, the current performance is limited. The destination must be visible from the robot initial position. Furthermore, possible feature loss due to consecutive feature update exists. An error case in SIFT matching due to consecutive update of the SIFT features is shown in Fig. 10. The left column shows the wide-angle and telephoto images at the first time step, while the right column shows the images at the 46-th frame. The feature window has slid upward, which caused a large error in the corresponding telephoto image. To avoid this kind of error, a memory of features over relatively long time interval may be helpful.



Fig. 10. Error in feature matching due to consecutive update of SIFT features.

V. CONCLUSIONS

In this paper, a destination for a human assisted mobile robot equipped with a multi-focal vision system is given by a user. The robot moves then toward a landmark at this destination, which is described by SIFT features in 2D wide-angle image. To achieve a natural landmark selection from the user perspective and an accurate feature tracking for a safe robot navigation, the image regions providing most number of SIFT features are provided to the user. The selected landmark is projected onto telephoto images to extend the sensing range. Optical flow technique is used to compute the corresponding points of the projected landmark on two telephoto images, in order to obtain an accurate relative position estimation between the landmark and the robot. A coordination strategy is realized, in which the camera with wide field of view is used for robot orientation control and the high-resolution camera is applied for robot forward motion control. The performance of our strategy is experimentally evaluated. This elaborate design and implementation of this coordination strategy is empirically validated to be advantageous for robot navigation in a human dominated environment.

ACKNOWLEDGMENTS

We would like to thank the other ACE-Team members (G. Lidoris, K. Klasing, A. Bauer, T. Zhang, Q. Mühlbauer, S. Sosnowski, F. Rohrmüller and D. Wollherr) for their excellent work designing and implementing the ACE platform. This work is supported in part within the DFG excellence initiative research cluster *Cognition for Technical Systems – CoTeSys*, see also www.cotesys.org.

REFERENCES

- [1] F. Schubert, T. Spexard, M. Hanheide and S. Wachsmuth. *Active Vision-based Localization For Robots In A Home-Tour Scenario*. In Proceedings of the 5th International Conference on Computer Vision Systems, 2007.
- [2] A. M. Arsenio. *Map Building from Human Computer Interaction*. IEEE CVPR Workshop on Real-Time Vision for Human Computer Interaction, 2004.
- [3] I. R. Nourbakhsh, C. Kunz, T. Willeke. *The robot museum robot installations: a five year experiment*. In Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2003.
- [4] R. D. Schraft, B. Graf, A. Traub, D. John. *A Mobile Robot Platform for Assistance and Entertainment*. In Industrial Robot Journal, Vol. 28, pp. 83-94, 2001.
- [5] M. Shiomi, T. Kanda, H. Ishiguro and N. Hagita. *Interactive humanoid robots for a science museum*. In Proc. of the 1st ACM SIGCHI-SIGART Conference on Human-robot interaction. pp. 305-312, 2006.
- [6] T. Xu, H. Wu, T. Zhang, K. Kühnlenz, and M. Buss. *Environment Adapted Active Multi-focal Vision System Using Kalman Filter for Object Detection*. In Proceedings of International Conference on Robotics and Automation, Kobe, Japan, 2009.
- [7] A. Bauer, K. Klasing, G. Lidoris, Q. Mühlbauer, F. Rohrmüller, S. Sosnowski, T. Xu, D. Wollherr, K. Kühnlenz and M. Buss. *The Autonomous City Explorer Project: Towards Natural Human-Robot Interaction in Urban Environments*. International Journal of Social Robotics 1, no. 2, 127-140, 2009.
- [8] E. Hayman, T. Thorhallsson and D. Murray. *Tracking while Zooming using Affine Transfer and Multifocal Tensors*. International Journal of Computer Vision, 51(1), 2003.
- [9] J. C. Fiala, R. Lumia, K. J. Roberts, and A. J. Wavering. *Triclops: A tool for studying active vision*. International Journal of Computer Vision, 12(2-3): 231-250, 1994.
- [10] M. Pellkofer and E.D. Dickmanns. *EMS-Vision: Gaze Control in Autonomous Vehicles*. In Proceedings of the IEEE Intelligent Vehicles Symposium 2000, Dearborn, USA, 2000.
- [11] T. Shibata, S. Vijayakumar, J. Conradt and S. Schaal. *Humanoid Oculomotor Control Based on Concepts of Computational Neuroscience*. In Proceedings of International Conference on Humanoid Robots, Waseda University, Japan, 2001.
- [12] N. D. Jankovic and M. D. Naish. *Developing a modular spherical vision system*. In Proceedings of International Conference on Robotics and Automation, 2005.
- [13] A. Ude, C. Gaskett and G. Cheng. *Foveated Vision Systems with two Cameras per Eye*. In Proceedings of International Conference on Robotics and Automation, 2006.
- [14] K. Kühnlenz, M. Bachmayer, and M. Buss. *A multi-focal high-performance vision system*. In Proceedings of International Conference on Robotics and Automation, 2006.
- [15] K. Kühnlenz. *Aspects of Multi-Focal Vision*. PhD Thesis, Institute of Automatic Control Engineering, Technische Universität München, 2006.
- [16] D. G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints*. International Journal of Computer Vision (60), no. 2, pp. 91-110, 2004.
- [17] M. A. Fischler, R. C. Bolles. *Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography*. Communications of the ACM, Vol 24, pp 381-395, 1981.
- [18] B. D. Lucas and T. Kanade. *An iterative image registration technique with an application to stereo vision*. In Proceedings of image understanding workshop, pp. 121-130, 1981.
- [19] J. Y. Bouguet. *Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm*. CS 223-B: Introduction to Computer Vision, Stanford University, 2004.