Real-Time Active Visual Tracking System

S. Ribaric, G. Adrinek, and S. Šegvic
Faculty of EE and Computing/ZEMRIS, Zagreb, Croatia
slobodan.ribaric@fer.hr

***Abstract* — The paper describes implementation of a real-time visual tracking system with an active camera. The system is intended for indoor human motion tracking. Real-time tracking is achieved using simple and fast motion detection procedures based on frame differencing and camera motion compensation. Results of on-line person tracking are presented.**

I. INTRODUCTION

Visual tracking is one of the most important fields of dynamic computer vision and it provides fundamental technologies to develop real world computer vision applications: human tracking and identification, intelligent transportation, traffic flow measurement and object tracking in smart rooms [1]-[5].

In this paper we describe a design of a real-time visual tracking system whose aim is to detect and track a person in indoor environments. According to the categorization of the target tracking systems [6], the developed system can be classified as:

i) The system for detection and tracking of a single object (person) in indoor scenes;

   Comment: Even for this simple task, real-time detection and tracking in real world environments requires complex image processing methods, as well as an object model for discriminating the foreground object from the background scene.

ii) The system can observe a wider area by changing the gazing direction using a camera, which is mounted on a pan/tilt gimbal;

   Comment: The system has features of smooth pursuit systems [7] or (in terms of active vision) systems with gaze stabilization [8]: object motion characteristics and mechanical camera dynamics are taken into account.

iii) The system has a single visual sensor – a CCD camera mounted on a fixed pan/tilt gimbal.

The paper is organized as follows: The next section presents related works in the field of real-time tracking. Section III gives the system overview and describes its main components. Techniques for camera motion estimation and control (used to provide smooth pursuit tracking), assumed characteristics of object motion, and procedures for motion detection and target localization are described in Section IV. Experimental tracking results obtained in a real indoor environment are given in Section V. Some conclusions and future research directions are given in Section VI.

II. RELATED WORK

There are many references related to the problem of tracking human motion, see e.g. [1]-[3], [9]-[13]. In this Section we give a brief overview of the previous work in related research directions and implementation details.

In general, tracking requires image segmentation, i.e. identification of the semantically meaningful components of a frame. The segmentation for dynamic-scene analysis is usually based on motion. The motion in the scene can be used to perform quick and efficient segmentation, distinguishing the moving foreground from the static background. For these reasons, segmentation based on motion is widely used as a base for many tracking methods. In our system we use a simple motion detection method based on a binary difference picture and size filtering [14]. This method (which is used for SCMO - Stationary Camera Moving Objects systems) has to be modified for MCMO - Moving Camera Moving Objects in such way that the compensation of motion introduced by camera movement has to be applied. There are several researchers, which have addressed this problem in a similar way. Murray and Basu [15] extract moving edges from image sequences taken by a rotating camera. Motion of the camera is compensated using transformations between frames based on camera parameters and known motion of the camera. Cai, Mitiche and Aggarwal [13] estimate motion (rotation) of the viewing system by matching line features between two frames. They use the fact that indoor environments usually contain lines that can be used for matching and that curved images of humans cannot influence the matching process significantly. People moving in the scene are detected using differences between camera motion compensated frames. Tracking based on moving edge detection and camera motion compensation in a sequence taken by a moving camera is also found in the work of Araki, Matsuoka, Yokoya and Takemura [16]. The ASSET-2 system [17] performs real-time tracking of road vehicles in a sequence taken by a moving camera. No calibration or knowledge of camera motion is needed. Motion estimation and tracking is based on detection and matching of image features - corners. Their method shows very good results even in the presence of arbitrary camera motion, but it is not applicable to our problem (tracking humans) because humans move quite differently than vehicles and do not contain many corner features. Instead, we have adopted their method for detection and matching of features used for compensation of camera motion. Real-time system performing active tracking is reported by Nordlund and Uhlin [18]. In the case when one moving object occupying only a small portion of the image, the object is found as the area that shows motion inconsistent with background motion. Background motion is calculated by finding one image velocity that is valid in the whole image using either affine velocity model or translational velocity model and the brightness constancy constraint. Another approach to segmentation for tracking systems is based on (adaptive) background modeling and background subtraction [19]-[21].

III. System Overview

The real-time tracking system functionality can be split into three basic phases: i) image acquisition and camera motion estimation; ii) object motion detection and localization; and iii) camera control.

Data acquisition is performed by a wide-angle CCD camera mounted on a fixed pan/tilt gimbal unit. Feature matching between two successive frames is used to estimate camera motion. Moving object is detected and localized by the following sequence of operations: i) thresholded frame difference; ii) morphological opening; iii) size filtering; iv) clustering. When the moving object is localized, commands are sent to the pan/tilt gimbal in order to perform a smooth pursuit.

IV. Tracking

A. Camera motion estimation

If we assume that pan and tilt angles between two successive frames are small enough, it is possible to find regions in these frames that contain almost the same view of the scene, i.e. common regions. Finding common regions of two successive frames is the key task of the system. For these regions, fast motion detection and object localization methods can be used. Common regions of two frames are found by determining translational displacement between them. This displacement is caused by the camera (ego) motion.

Based on the assumption that the moving object occupies smaller portion of the scene, camera motion is estimated as the dominant motion - the motion of the background (relative to the camera). Dominant motion is found as the most frequent motion vector in the image. Motion vectors are calculated for distinct features in the image – corners detected using the SUSAN corner detector [22]. These motion vectors are a result of a feature matching process, similar to one used in [17]. Fig. 1 shows an example of the feature matching process and dominant motion calculation.

(a) (b)

Figure 1. Successive frames with marked detected corner features and feature velocities obtained from matching process (the calculated dominant feature velocity is shown by the bold vector in (b)).

Comment: Another possible approach for estimating the camera motion would be to read pan/tilt positions directly from the gimbal unit [15]. Such an approach would provide increased simplicity at the cost of being less general since it could not be used in applications where the exact orientation of the camera is not known. The main shortcoming of that approach is the difficulty of obtaining position readings which are synchronized with the current image being processed. Although in theory one could compensate the latencies of digital video and the slow communication with the pan/tilt gimbal, that is a hard task in practice, which was the main reason for estimating the camera motion by a more general approach.

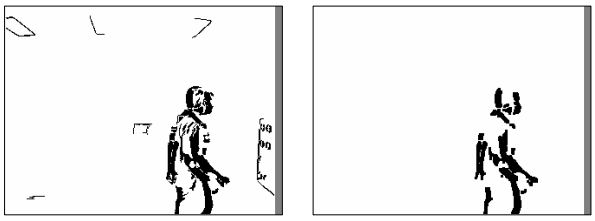
B. Object motion detection and localization

When the displacement between the two frames is known, i.e. their common regions are determined, we perform thresholded differencing of common regions in order to detect changes in the scene caused by object motion:

$$DP_{i,j}(x,y) = \begin{cases} 1 & \text{if } |F(x,y,i) - F(x+d,y,j)| > T \\ 0 & \text{otherwise} \end{cases}$$

where $F(x,y,i)$ is the pixel value at location (x,y) in frame i, d is the displacement between two frames, and T is a threshold. Result of this process is the difference picture DP, a binary map of pixels whose values change significantly between the two frames.

Some entries in the difference picture are a result of noise, and they need to be filtered out. We used simple size filtering [14]. This filter eliminates noise, but leaves some "phantom" regions (Figure 2a). These erroneous regions are a result of approximations made regarding common regions of frames (approximations are less valid as the angles between frames are larger). Common regions of two frames do not contain exactly the same view of the scene because each frame is taken from a different angle. Lens distortions also cause such errors. As can be seen in Fig. 2a, erroneous regions are very thin. Like Murray and Basu [15] we also used this property to eliminate such regions using morphological opening (structuring element size 3×3) which eliminates most of the erroneous regions and leaves regions caused by object motion (Fig. 2b).



(a) (b)

Figure 2. Difference picture of common regions of frames from Figure 3 (a) and a result of applying morphological opening to the difference picture (b). Gray areas represent the area of the current frame not found in the previous frame.

The remaining regions in difference picture are (hopefully) only a result of object motion in the scene. One moving object usually causes more than one region in the difference picture to appear. Also, regions in the difference picture can be a result of different moving objects in the scene. For that reason we perform spatial clustering of regions in the difference picture. A cluster that consists of a largest number of regions is considered as the object. We chose this criteria having in mind that we want to track humans which, when moving, tend to create a larger number of regions in the difference picture, unlike rigid moving objects. Examples of target localization are presented in the experiments section.

C. Camera control

When the object is localized, the camera should be directed towards it. We use a simple control strategy

which minimizes the amount of communication with the pan/tilt unit and enables smooth object pursuit even when the target is not precisely localized from frame to frame. Pan axis angular velocity is set on a value proportional to the difference between the center of the image and the center of the object's enclosing rectangle: $w = C \cdot d$, where C is a constant which determines the speed of tracking and d is the difference between the center of the image and the center of the target. When d is small enough, angular velocity is not changed. This method makes tracking possible even when the field of view angle is not known.

Comment: We used only pan axis for tracking because, in most indoor setups, it offers sufficient freedom for following horizontal human movement.

V. Experimental Results

During the development of the system, many experiments were performed in order to tune the parameters of the system. Here we present only the experiments closely related with the performance of the system - tracking speed and the stability of tracking (ability to recover from errors during the tracking).

A. Experimental setup

Experiments were performed using a digital camera with a wide angle (64°) lens, connected to the *IEEE 1394* bus. The pan/tilt unit can speed up over 300°/second but uses slow communication interface (RS-232). Processing is performed on a dual processor computer with SPEC int_rate2000 of 12.1. One processor was used for image acquisition and the other for actual processing. In order to enable real-time processing, the original image format of 640×480 was programmatically reduced to 320×240.

B. Performance

Using this setup, image processing, including image acquisition, camera motion compensation, change detection and target localization, runs at 25 Hz. When pan velocity command is issued, which lasts up to 17.5 ms, the performance falls to around 17 Hz. Because pan velocity command is not issued for each frame, the average processing rate is around 20 Hz.

The maximum tracking speed (angular velocity by which the camera is directed towards the target) of 50°/s, for which the system shows good tracking stability, was determined experimentally.

C. Tracking

Many experiments were performed, in different indoor environments, with different types of movement and with different people, and in different lighting conditions. The experiments showed that the system successfully tracks a single walking or running person. Examples showing tracking of a person walking and running are shown in Fig. 3 and Fig. 4, respectively (both figures contain images taken in online experiments).

VI. Conclusions and Future Work

We have presented an implementation of a real-time human motion tracking system. By combining the method used with static camera with camera ego-motion compensation we managed to design a system capable of efficacious person tracking in indoor enviroment. Ego-motion is determined using corner feature matching. The system was tested on-line under various lighting conditions. The experimental results show robustnes of the system applied to human motion tracking with average processing rate of 20 frames/s and maximum tracking speed of 50°/s. Due to object motion characteristics and equipment limitations (comunication with pan/tilt gimbal), the system uses only pan axis control for smooth pursuit.

Future work will be directed towards dealing with both vertical and horizontal components of object motion, introduction of zoom control and multiple sensor configurations.

Figure 3. A walking person sequence (every 15th frame is shown): tracking fails only when the object is in front of the background with similar color.

Figure 4. A running person sequence (every 15th frame is shown): tracking fails only when the object is in front of the background with similar color..

REFERENCES

[1] C. Wren, A. Azarbayeani, T. Darrell, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 1997, vol. 19, no. 7, pp. 780 -785.

[2] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Real-Time Surveillance of People and Their Activities", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, August 2000, pp. 809-830.

[3] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomato, O. Hasegawa, "A System for Video Surveillance and Monitoring: VSAM Final Report", *Technical report CMU-RI-TR-00-12*, Robotics Institute, Carnegie Mellon University, May 2000.

[4] T. S. Lee, E.M. Lee, H.T. Pak, Y.K. Kwang, S.S. Lim, J.H. Beak, B. W. Hwang, "Implementation of Traffic Flow Measuring Algorithm Using Real-Time Dynamic Image Processing", in Computer Vision Systems (J. L. Crowley et al., eds), LNCS 2626, Springer, 2003, pp. 78-87.

[5] D. R. Karuppiah, Z. Zhu, P. Shenoy, E. M. Rieseman, " A Fault-Tolerant Distributed Vision system Architecture for Object tracking in a Smart Room", in Proceed. of the International Workshop on Computer Vision Systems, Vancouver, Canada , 2001, pp. 201-219.

[6] T. Matusuyama, N. Ukita, "Real-Time Multi-Target Tracking by Cooperative Distributed Vision System", Proceed. of the IEEE 90, 2002, pp. 1136-1150.

[7] K. Daniilidis, C. Krauss, M. Hansen, and G. Sommer, "Real-Time Tracking of Moving Objects with an Active Camera", *Real-Time Imaging* 4, 1998, pp. 3-20.

[8] J. Hynoski, and H. R. Wu, "Active vision – a survey of the field and research directions", *Technical Report 95-04*, Monash University, Faculty of Computing and Information Technology, Department of Robotics and Digital Technology, Clayton, Australia, 1995.

[9] Q. Cai, J. Aggarwal, "Tracking Human Motion in Structured Environment Using a Distributed-camera System", IEEE Trans. of the PAMI, 21, 1999, pp. 1241-1247.

[10] A. Nakazawa, H. Kato, S. Inokuchi, "Human tracking Using Distributed Vision Systems", in *Proceed. of the Int. Conference on Pattern Recognition*, Vol. I, Brisbane, Australia, IEEE (1998), pp. 595-596.

[11] S. Khan, O. Javed, Z. Rasheed, M. Shah, "Human tracking in multiple cameras", Proceed. of the Int. Conference on Computer Vision, Vancouver, Canada, IEEE (2001), pp.331-336.

[12] S. L. Dockstader, A. Tekalp, "Multiple Camera Tracking of Interacting and Occluded Human Motion", *Proceed. of the IEEE*, 89, 2001, pp. 1441-1455.

[13] Q. Cai, A. Mitiche, and J. K. Aggarwal, "Tracking Human Motion in an Indoor Environment", *International Conference on Image Processing*, 1995, vol. 1, pp. 215 – 218.

[14] R. Jain, R. Kasturi, B. G. Schunck, "Machine Vision", McGraw-Hill, 1995.

[15] D. Murray and A. Basu, "Motion Tracking with an Active Camera", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no 5, 1994, pp. 449-459.

[16] S. Araki, T. Matsuoka, N. Yokoya, and H. Takemura, "Real-Time Tracking of Multiple Moving Object Contours in a Moving Camera Image Sequence", *IEICE Tranactions on Information and Systems*, Vol.E83-D, No. 7, July 2000.

[17] S. M. Smith and J. M. Brady, "ASSET-2: Real-time motion segmentation and shape tracking", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, 1995, pp. 814-820.

[18] P. Nordlund and T. Uhlin, "Closing the Loop: Detection and Pursuit of a Moving Object by a Moving Observer", *Technical Report CVAP-175-95-7-173*, Computational Vision and Active Perception Laboratory, Royal Institute of Technology, 1995, Stockholm, Sweden.

[19] C. Eveland, K. Konolige, R. Bolles, "Background Modeling for Segmentation of video-rate Stereo Sequences", *Proceed. of the CVPR 98*, 1998, pp. 266-271.

[20] Q.Z. Wu, B.S. Jeng, "Background Subtraction Based on Logarithmic Intensities", *Pattern Recognition Letters* 23, 2002, pp. 1529-1536.

[21] S. Stauffer, W.E.L. Grimson, "Adaptive Background Mixture Models for Real-time Tracking", *Proceed. of the CVPR 99*, June 1999, Vol. 2, pp. 246-252.

[22] S. M. Smith, J. M. Brady, "SUSAN – A New Approach to Low Level Image Processing", *Int. Journal of Computer Vision*, 23(1), May 1997, pp. 45-78

[23] D. M. Gavrila, "The Visual Analysis of Human Movement: A Survey", *Computer Vision and Image Understanding*, 73(1), January 1999, pp. 82-98