8—20

3D Modelbased Detection of People in Monocular Video Sequences taken in Trainstations

Stefan Huwer, Heinrich Niemann
FORWISS - Knowledge Processing Research Group
Bavarian Research Center for Knowledge-Based Systems
Am Weichselgarten 7
D-91058 Erlangen
{Stefan.Huwer}@forwiss.de
Phone: +49-9131-691139 FAX: +49-9131-691185

Abstract

This article focuses on a new approach for three-dimensional people detection in monocular image sequences taken from stationary cameras with fixed focal lengths. The main contribution of the new algorithm is the combination of the following four parts. An abstract articulated geometrical person model, a scene model including a platform and a camera model, a hypothesis generator for people positions based on the CONDENSATION algorithm, and a fast clustering approach for the people detection. This approach allows the robust detection of several persons on the platform in realtime. This approach demonstrates that even with monocular image processing algorithms it is possible to detect people in realtime and to obtain information on their three-dimensional locations.

1 Introduction

The work presented here is part of the research project "Intelligent Train Plattform' which is concerned with automatic visual surveillance of subway train stations by image processing methods. In order to increase passenger safety and minimize train stop time, the end of the passenger change must be detected accurately and reliably. Two independently operating modules, namely surveillance of the train and surveillance of the passengers on the platform, yield three-dimensional information about the scene, which allows the detection of the end of the passenger change. The realworld character of the application demands special needs: The modules should run in "realtime" and since the number of cameras has to be minimized in order to safe expenses at each train station, the modules must operate successfully on monocular image sequences. The method presented allows the monocular detection of people in realtime. However this approach is not applicable to the detection of mostly occluded persons or crowds (see [FNW98] for a method that allows the detection of crowds of people).

The person detection module can be decomposed into three submodules (see figure 1), the image-sequence preprocessor, the modelling part and the detection module (the hypothesis generator), which are explained in the following.

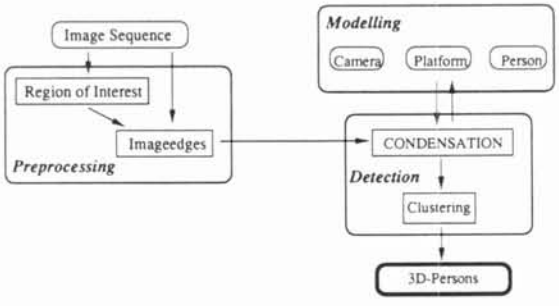


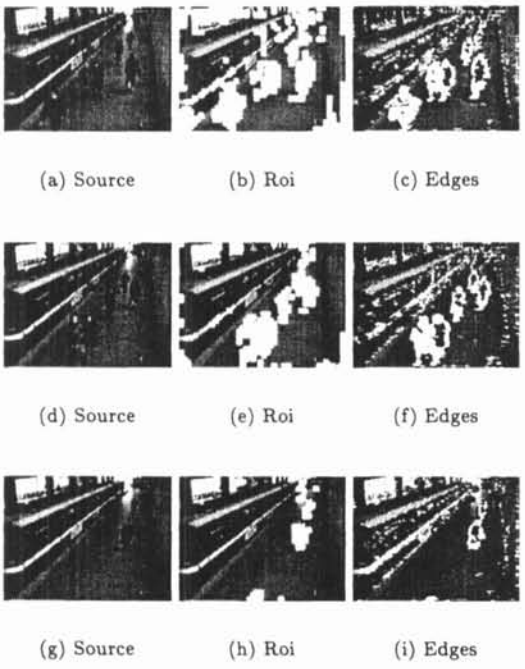
Figure 1: Program Structure

2 Preprocessing

An adaptive change detector [HN00] supplies the edge detector with a region of interest in which edges are to be extracted. The adaptive change detector combines a temporal difference method with an adaptive background model subtraction scheme. When changes in illumination occur, the background model is automatically adapted to suit the new conditions. For the adaptation of the background model a method is exploited, which avoids reinforcement of adaptation errors by performing the adaptation solely on those background regions that were detected by the temporal difference method rather than using the regions resulting from both, the temporal difference method and the background subtraction method.

A Sobel operator has been found to be sufficiently reliable for the edge detection. Based on the internal and external camera calibration data, the morphological operations dilation and erosion are applied to the edges in order to scale the thickness of edges according to their distance from the camera. Figure 2 shows some result images of the detection of

foreground.



(a) Source (b) Roi (c) Edges

(d) Source (e) Roi (f) Edges

(g) Source (h) Roi (i) Edges

Figure 2: Examples for preprocessing

3 Models

The modelling part of the system consists of the camera model, the scene model and the person model.

3.1 Camera Model

The camera model provides the link between the image sequence and the scene modelled in three dimensions and allows the projection of the person models into the image domain. The camera is modelled as a pinhole camera with radial distortion (as suggested in [LT88]).

Using this projective cameramodel, it is possible to transform any given 3D point from world coordinates into pixel coordinates and to transform a 2D pixel into a world ray. Thus an instantiated geometrical 3D model can be transformed into a 2D image representation.

3.2 Person Model

Two different person models have been used for detection and tracking, with similar results. Both models can be reduced to the same abstract model, which is given by 12 joints and 36 degrees of freedom, including the position and orientation of the model in the scene. The first model is the commercially available person model RAMSIS[1] which amongst other things has been developped for the analysis of ergonomics. It comes with a variety of functions for pose estimation and pose analysis but it is not suitable for realtime applications because of its computational complexity. The second person model GEOM is a much simpler 3D geometrical person model which allows the fast rendering of a given model state into a 2D representation. A rendered view of each model can be seen in figure 3. A two-dimensional shape of a person model positioned and oriented in 3D can be extracted via the projective transformation. For a real-time implementation the shape generation uses the rendering machine OpenGL. Examples of shapes of the model can be seen in figure 4 on the right hand side.

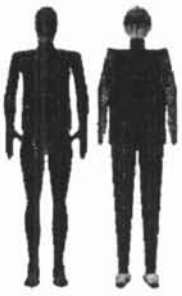


Figure 3: Person Models: RAMSIS (left), GEOM (right)

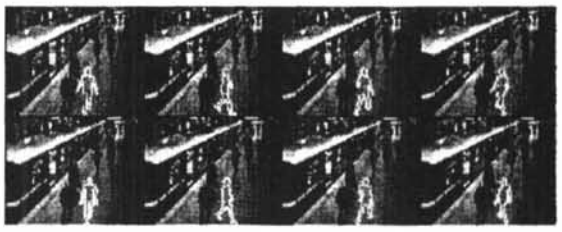


Figure 4: Rendered Person Models: Upper row GEOM, Lower row RAMSIS

3.3 Scene Model

The scene model is used in order to determine 3D view rays from 2D pixels by simple linear algebra. For a given pixel a view ray can be computed in 3D world coordinates. If, for example, the view ray intersects the pixel of the foot of a person in the image domain, the 3D position of this person is given at intersection of this ray with the scene model. This is subject to the condition that the person stands on its feet and the scene model is a surface model. In the application "Intelligent Train Platform" the scene model is given by a 3D plane with boundaries of the platform in 3D coordinates.

The next paragraph describes how hypotheses about locations of people are determined by uti-

[1]TECMATH GmbH, Kaiserslautern, Germany

lizing the results of the preprocessing module and the knowledge of the modelling module.

4 People Detection

The detection algorithm is based on the Condensation algorithm for the propagation of conditional densities[2]. In order to use the Condensation algorithm, a state space must be given which describes the possible states of persons to be detected. The state space $\Sigma \subset \mathbb{R}^8$ can be defined as the 3D positions of of persons on the platform, an angle of orientation of the person which describes the rotation of the person model around its length axis, as well as the corresponding four first moments, which represent the vector of change , for the internal prediction of the position of the hypothesis into the next image frame:

$$\Sigma := \left\{ \begin{pmatrix} x \\ y \\ z \\ \alpha \\ \dot{x} \\ \dot{y} \\ \dot{z} \\ \dot{\alpha} \end{pmatrix} \in \mathbb{R}^8 \;\middle|\; \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \text{“Platform”} \text{ and } \alpha \in [0, ..., 2\pi[\right\}$$

For each image the algorithm generates a fixed number of state hypotheses. A hypothesis is generated either at random or is the result of a statistical propagation of a *good* hypothesiss from the last image. In a second step the corresponding shape of the model is extracted for each of the hypotheses and its fit to the computed image edges is evaluated. As a measure of fit the relative overlap between the model shape and the image edges is computed:

$$Overlap = \frac{\text{Area}(Modelshape \bigcap \text{Imageedges})}{\text{Area}(Modelshape)}$$

Overlap is the quality measure chosen for each hypothesis and its modelshape. As the condensation algorithm usually yields several hypotheses in the vicinity of each person, a clustering approach was chosen to select single persons from the hypotheses. Clustering is carried out by projecting the quality values of all hypotheses onto a subsampled ground plane of the scene model (the platform), where a tile of the plane represents a 0.5[m] times 0.5[m] square region of the platform. This subsampled plane can be treated as a two-dimensional image which allows the detection of clusters with fast image processing procedures. After smoothing the cluster image , the local maximal points of the image represent the initial points of the clusters. For each initial point the best hypothesis within a certain neighborhood gives the cluster center. Those cluster centers yield the final hypotheses which are the inputs for the tracking module. This detection process is shown in figure 5.

(a) Source (b) Clusters (c) Detected Persons

Figure 5: Hypotheses Clustering

In order to eliminate erroneous noise in hypotheses the final hypotheses are evaluated for a second time on the basis of their *Overlap* which must exceed a certain threshold. The set of hypotheses $\Lambda := \{\lambda_1, ..., \lambda_n\} \subset \Sigma$ gives the measurements for the next module, the multi Kalman tracker. Examples for detected persons are shown in figure 6, where the shapes, corresponding to the persons state, are rendered into the input image.

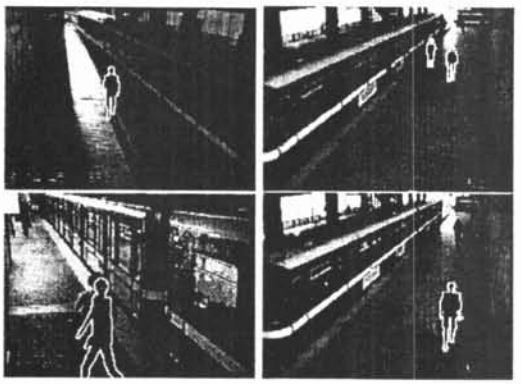


Figure 6: Examples for person hypotheses

It can be noted that not every person is detected at every time step. But as long as each person is detected in a reasonable time, the tracking algorithm is able to handle it.

5 Conclusions

The performance of the algorithm depends on the number of hypotheses chosen, which have to be evaluated for each image. A good recognition rate was achieved with 200 hypotheses, which resulted in a processing rate of approximately 10 frames per second, running on a SUN Ultra 10 with an image size of 384x288 pixel. Results of the algorithm are shown in figure 7, where the model shapes are shown at the detected positions.

References

[FNW98] D. Faulhaber, H. Niemann, and P. Weierich. Detection of crowds of people by use of wavelet features and parameter free statistical models. In

[2] An explanation of the Condensation algorithm can be found in [Isa98]

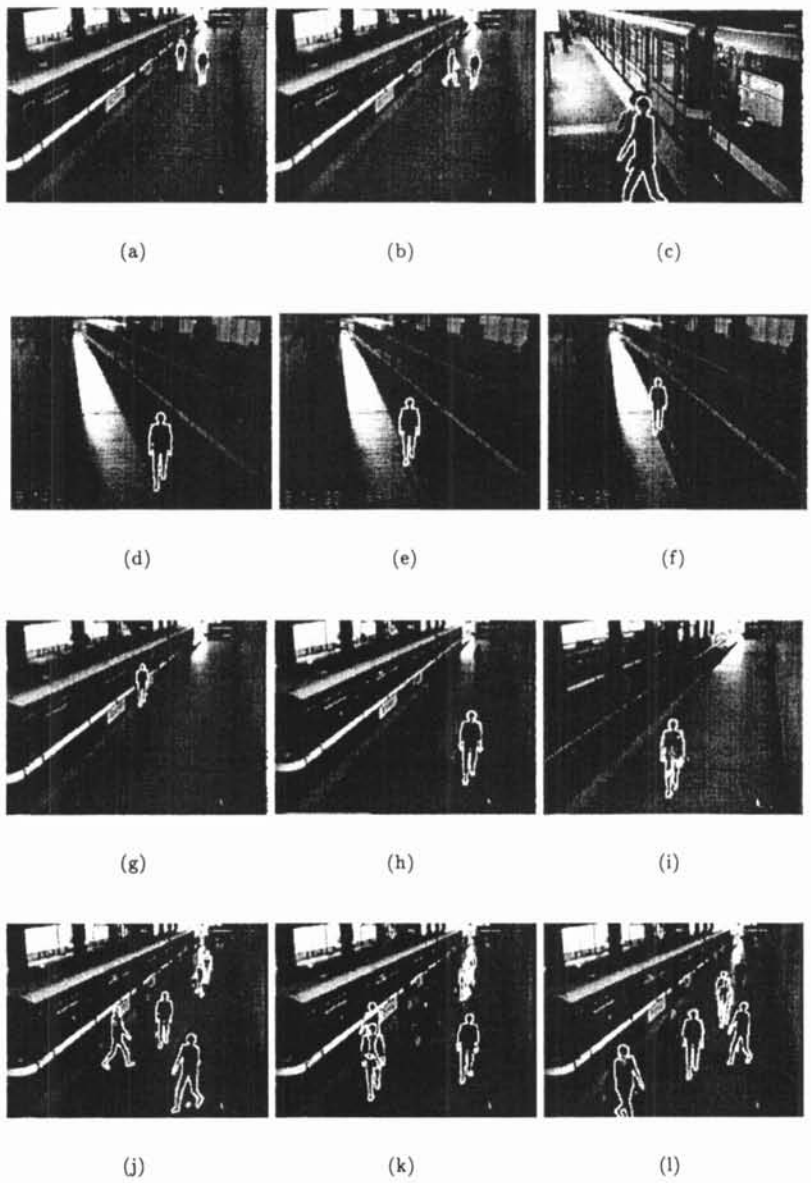


Figure 7: Detected Persons

IAPR Workshop on Machine Vision and Applications, Tokyo, Japan, 1998.

[HN00] S. Huwer and H. Niemann. Adaptive change detection for real-time surveillance applications. In *Third IEEE International Workshop on Visual Surveillance*, pages 37–45, Dublin, July 2000. IEEE, IEEE Computer Society, Los Alamitos.

[Isa98] M. A. Isard. *Visual Motion Analysis by Probabilistic Propagation of Conditional Density*. PhD thesis, Robotics Research Group, Department of Engineering Science, University Oxford, September 1998.

[LT88] R.K. Lenz and R.Y. Tsai. Techniques for calibration of the scale factor and image center for high accuracy 3-d machine vision metrology. *IEEE Transactions on Pattern Analysis and Machine Intelligene*, 10:713–720, 1988.