Shape-based pedestrian detection and localization

M. Bertozzi*, A. Broggi*, R. Chapuis†, F. Chausse†, A. Fascioli*, and A. Tibaldi*
*Dip. di Ingegneria dell'Informazione
Università di Parma – ITALY
†LASMEA UMR 6602 UBP/CNRS
Université de Clermont-Ferrand – FRANCE

***Abstract*— This work presents a vision-based system for detecting and localizing pedestrians in road environments by means of a statistical technique.**

Initially, attentive vision techniques relying on the search for specific characteristics of pedestrians such as vertical symmetry and strong presence of edges, allow to select interesting regions likely to contain pedestrians. These regions are then used to estimate the localization of pedestrians using a Kalman filter estimator.

I. INTRODUCTION

The pedestrians detection is an essential functionality for intelligent vehicles, since avoiding crashes with pedestrians is a requisite for aiding the driver in urban environments.

Vision-based pedestrian detection in outdoor scenes is a challenging task even in the case of a stationary camera. In fact, pedestrians usually wear different clothes with various colors that, sometimes, are barely distinguishable from the background (this is particularly true when processing grey-level images). Moreover, pedestrians can wear or carry items like hats, bags, umbrellas, and many others, which give a broad variability to their shape.

When the vision system is installed on-board of a moving vehicle additional problems must be faced, since the observer's ego-motion entails additional motion in the background and changes in the illumination conditions. In addition, since Pedestrian Detection is more likely to be of use in a urban environment, also the presence of a complex background (including buildings, moving or parked cars, cycles, road signs, signals...) must be taken into account.

Widely used approaches for addressing vision-based Pedestrian Detection are: the search of specific patterns or textures [1], stereo vision [2]–[4], shape detection [5]–[7], motion detection [8]–[10], neural networks [11], [12]. The great part of the research groups use a combination of two or more of these approaches [2], [13], [14]. Anyway, only a few of these systems have already proved their efficacy in applications for intelligent vehicles.

This work presents the first results of a new localization and association rule specifically designed to follow the detection process previously developed [15].

In this work the strong vertical symmetry of the human shape is exploited to determine specific regions of interest which are likely to contain pedestrians. This method allows the identification of pedestrians in various poses, positions and clothing, and is not limited to moving people. In order to improve the reliability of the system and as preliminary work for pedestrain tracking, a pedestrian localization step has been added. Pedestrian localization iteratively computes the position of pedestrians in the 3D world. It has been conceived to be used for a tracking system.

This paper is organized as follows. Section 2 introduces the structure of the algorithm. Section 3 describes the detection module, section 4 presents the localization procedure. Section 5 ends the paper with some final remarks.

II. ALGORITHM STRUCTURE

Figure 1 shows the algorithm structure. As a first processing step, attentive vision techniques are applied to concentrate the analysis on specific regions of interest only. In fact, the aim of the low-level part of the processing is the focusing on potential candidate areas to be further examined at a higher-level stage in a following phase.

The areas considered as candidate are rectangular bounding boxes which:

- have a size in pixels falling in a specific range. This range is computed from the knowledge of the intrinsic parameters of the vision system (angular aperture and resolution) and from allowed size and distance of pedestrians;
- enclose a portion of the image which exhibits the low-level features that characterize the presence of a pedestrian, i. e. a strong vertical symmetry and a high density of vertical edges.

A stereo refinement is used to refine the computed bounding boxes. The other image is searched for the same detected object and a triangulation is used to determine the distance.

Moreover, since other objects than pedestrians feature high symmetrical content, a set of filters is used to remove objects like poles, trees....

The forward loop process ends by estimating the pedestrian position in the road scene. This stage uses an internal model (the *scene bounding box*) that allows to take into account the possible bad fitting of the detected bounding box with respect to the real pedestrian shape.

Beside the obvious usefulness of a pedestrian localization functionality for a driver assistance system (for example, in order to know which pedestrian is the more dangerous and to focus the perception on him); in addition, as shown on figure 1 (dotted lines) localization can also be use to foresee the future position of the bounding boxes (i.e. to track the pedestrians). The tracking can be used to directly act onto

0-7803-8125-4/03/$17.00

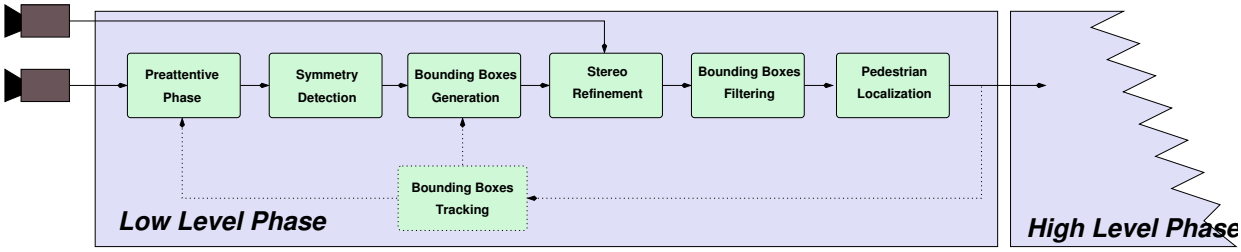

Fig. 1: The algorithm architecture.

the detection stages in order to improve the reliability of the system. Currently, the loop has not been yet closed. The full tracking system is under developement.

III. Pedestrian detection

A. Search area

In the first phase a search for pedestrians candidates is performed. Thanks to the knowledge of the system's extrinsic parameters together with a flat scene assumption, this search is limited to a reduced portion of the image. Besides the obvious advantage of avoiding false detections in wrong areas, the processing of a reduced search area only, reduces the computational time. The analysis is not limited to a target featuring a fixed size or a given distance, but a range for each parameter is in fact considered. The introduction of these ranges generates two further degrees of freedom in the size and position of the bounding boxes. In other words, the search area is enlarged to accommodate all possible combinations of height, width, and distance for pedestrians.

B. Symmetry detection

Since pedestrians are generally symmetric, the processing is based on the analysis of symmetries. Shadows and other road textures do not influence the search. The analysis proceeds in this way: the columns of the image are considered as possible symmetry axes for bounding boxes. For each symmetry axis different bounding boxes are evaluated scanning a specific range of distances from the camera (the distance determines the position of the bounding box base) and a reasonable range of heights and widths for a pedestrian (the corresponding bounding box size can be computed through the calibration).

However, not all the possible symmetry axes are considered: since edges are chosen as discriminant in most of the following analysis, a pre-attentive filter is applied, aimed at the selection of the areas with a high density of edges. Axes centered on regions which contain a number of edges lower than the average value are dropped.

For each of the remaining axes the best candidate area is selected among the bounding boxes which share that symmetry axis, while having different position (base) and size (height and width). Vertical symmetry has been chosen as a main distinctive feature for pedestrians. Symmetry edge maps, e. g. the Generalized Symmetry Transform (GST) [16], have already been proposed as methods to locate interest points in the image prior to any segmentation or extraction of context-dependent information. Unfortunately, these methods are generally computationally expensive. Alternatively, two different symmetry measures are performed: one on the gray-level values and one on the horizontal gradient values. The selection of the best bounding box is based on maximizing a linear combination of the two symmetry measures, masked by the density of edges in the box.

C. Bounding boxes generation

An adjustment of the bounding boxes' size is yet needed. In fact, when comparing the gray-level symmetry of different bounding boxes centered on the same axis, larger boxes tend to overcome smaller ones since pedestrians are generally surrounded by homogeneous areas such as concrete underneath or the sky above. Therefore, the bounding box which presents the maximum symmetry tends to be larger than the object it contains because it includes uniform regions. For this reason, for each selected symmetry axis, the exact height and width of the best bounding box are actually taken as those possessed by the box which maximizes a new function among the ones having the same axis. This function is computed as the product of the symmetry of vertical edges and density of vertical edges only. Figure 2 summarizes the overall candidate generation process.

The result of this step is a first list of candidate bounding boxes that contains potential pedestrians.

D. Stereo refinement

The distance of the potential pedestrians can be computed using the knowledge of the camera calibration and the assumption of a flat scene. Unfortunately, the computed values are greatly affected by a wrong detection of the lower part of pedestrians. In order to refine this measurement, which is of importance for discriminating amongst obstacles and actual pedestrians, a refinement phase is mandatory.

A simple stereo technique is used: for each bounding box in this list, starting from a rough estimation of the distance, a portion of the other image is searched for areas which exhibit a content similar to the one included in the bounding box by means of a correlation measure. The correlation formula used for matching left a_i and right b_i pixels is:

$$\chi = \frac{(\sum a_i b_i)^2}{(\sum {a_i}^2)(\sum {b_i}^2)}$$

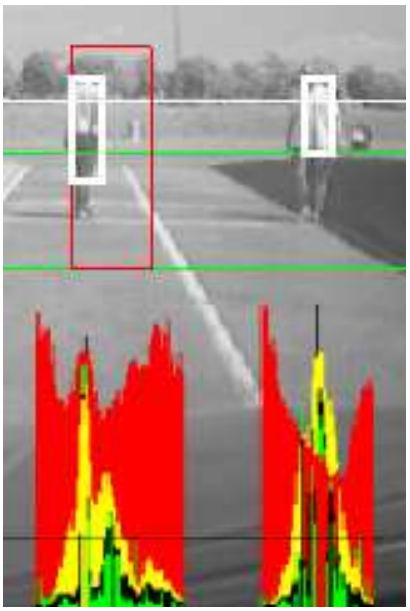

Fig. 2: Bounding boxes generation phase: the two horizontal green lines represent the search area; the symmetry histograms for grey-levels (red), vertical edges (green), the vertical edges density (yellow), and their combination (black) are shown in the bottom part of the image.

Once the correspondence between the bounding box located in the left image and its counterpart in the right image has been found, a triangulation is be used to determine the distance to the vision system. Therefore, a refinement of the bounding box base can take place, based on calibration and perspective constraints. More precisely, the knowledge of the camera orientation with respect to the ground and the road slope can provide information about the position of the point of contact of the human shape with the ground. This knowledge is used to stretch the bottom of the bounding box till it reaches the ground and frames the entire shape of the pedestrian and the technique is robust in the sense that even if the background is different from one image to another, the distance is correctly evaluated and the base exactly refined for all observed cases (see figure 3).

E. Bounding boxes filtering

Symmetrical objects other than pedestrians may happen to be detected as well. In order to get rid of such false positives a number of filters have been devised which rely on the analysis of the distribution of edges within the bounding box and on segmentation and classification of the box region. These filters, which are still under development, show promising results regarding the elimination of both artifacts (such as poles, road signs, buildings, and other road infrastructures) and symmetrical areas given by a uniform portion of the background between two foreground objects with similar lateral borders.

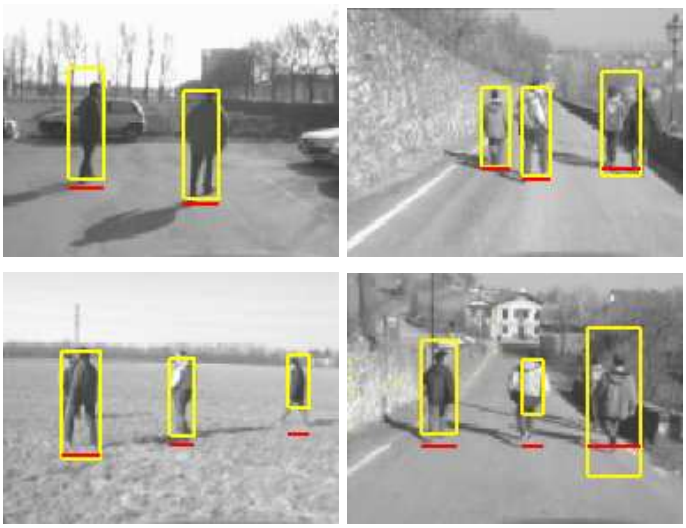

Fig. 3: Stereo refinement: the yellow bounding box is generated during the symmetry detection, the red line represents the stereo refinement of the box's bottoms.

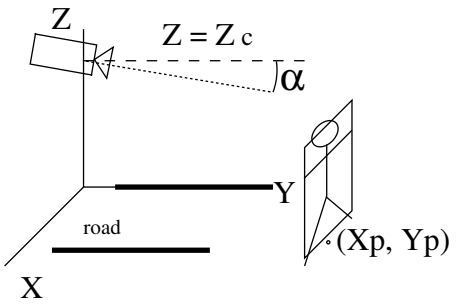

Fig. 4: Coordinates of a pedestrian.

IV. Pedestrian Localization

This section describes how the position of the pedestrian in the scene is estimated.

A. Pedestrian model and coordinate systems

A pedestrian is considered as being situated on a flat road and its position is defined by the coordinates $(X_p, Y_p, 0)$ corresponding to the intersection of its vertical axis with the ground (i.e. the contact point). Figure 4 shows the coordinate system in which the contact point is defined according to the road sides.

This scene is observed by a camera situated at height Z_c with a tilt angle denoted by α as shown on figure 4.

B. Pedestrian observation

We take here into account that the detection provides image bounding boxes including a pedestrian shape. Unfortunately, sometimes a bounding box does not fit exactly the pedestrian; in particular the height of the bounding box is smaller than the pedestrian shape. This particularity is of great importance for the localization since the observed height influences directly the estimated distance from the camera to the pedestrian.

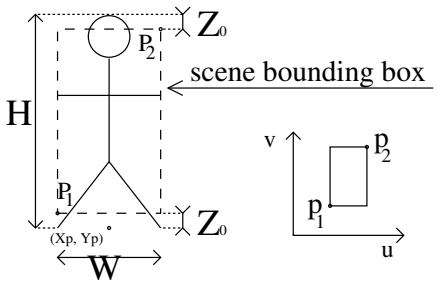

Fig. 5: Correspondent reference points projection.

The position estimation process must take explicitly this into account in order to provide a more accurate result.

To solve this problem, a scene bounding box is defined (figure 5). It is supposed to correspond to the image bounding box provided by the detection. A scene bounding box must correspond to a pedestrian. So its is defined by a height H and by a width W. These last parameters are considered as random variables described by a normal law. Their average and standard deviation are chosen to be realistic enough to represent a human shape: a pedestrian is supposed to be a person whose average height is $H = 1.70\ m$ with a standard deviation $\sigma_H = 0.1\ m$, and whose width is strongly correlated to the height $W = k\ H$ with *a priori* $k = 0.3$ which correspond to a realistic width/height ratio, having standard deviation $\sigma_W = 0.1\ m$.

Z_0 represents half of the height difference between the scene bounding box and the real pedestrian height. Its average value is chosen null with a standard deviation of $\sigma_{Z_0} = 0.1m$.

The next operation consists in establishing the relationship between the scene bounding box and the pedestrian position parameters. The coordinates of the corners P_1 and P_2 of the scene bounding box are obtained from the position X_p and Y_p of the pedestrian and from H, W and Z_0 and are expressed in a coordinate system linked to the camera and according to classical small angle approximation (tilt angle α is low):

$$P_1 = \begin{pmatrix} X_1 \\ Y_1 \\ Z_1 \end{pmatrix} \quad P_2 = \begin{pmatrix} X_2 \\ Y_2 \\ Z_2 \end{pmatrix} \text{ with} \tag{1}$$

$$\begin{cases} X_1 = X_P - \frac{W}{2} \\ Y_1 = Y_P - \alpha(Z_0 - Z_c) \\ Z_1 = \alpha Y_P + (Z_0 - Z_c) \end{cases} \text{and} \begin{cases} X_2 = X_P + \frac{W}{2} \\ Y_2 = Y_P - \alpha(H - Z_0 - Z_c) \\ Z_2 = \alpha Y_P + (H - Z_0 - Z_c) \end{cases} \tag{2}$$

Next, to make the link between scene and image, a perspective projection is considered according to camera calibration parameters e_u and e_v supposed to be known. It makes it possible to establish the relationship between the image coordinates of the corners of the image bounding box ($p_1 = (u_1, v_1)^T$, $p_2 = (u_2, v_2)^T$) and the scene coordinates of the scene bounding box provided that image escape point is settled in (u, v) origin of figure 5:

$$\begin{cases} u_1 = e_u \frac{X_1}{Y_1} \\ v_1 = e_v \frac{Z_1}{Y_1} \\ u_2 = e_u \frac{X_2}{Y_2} \\ v_2 = e_v \frac{Z_2}{Y_2} \end{cases} \tag{3}$$

Equation system (3) is equivalent to :

$$\begin{cases} u_1 Y_P - \underline{\alpha u_1 (Z_0 - Z_c)} = e_u (X_P - \frac{W_{\text{box}}}{2}) \\ v_1 Y_P - \underline{\alpha v_1 (Z_0 - Z_c)} = \underline{e_v \alpha Y_P} + e_v (Z_0 - Z_c) \\ u_2 Y_P - \underline{\alpha u_2 (H_{\text{box}} - Z_0 - Z_c)} = e_u (X_P + \frac{W_{\text{box}}}{2}) \\ v_2 Y_P - \underline{\alpha v_2 (H_{\text{box}} - Z_0 - Z_c)} = \underline{e_v \alpha Y_P} + e_v (H_{\text{box}} - Z_0 - Z_c) \end{cases} \tag{4}$$

Low α allows to neglect all the underlined parts of equation 4. Then the system becomes linear according to the pedestrian position coordinates X_p and Y_p and yields to the observation equation :

$$\begin{pmatrix} -e_v(Z_0 - Z_c) \\ e_u \frac{W}{2} \\ -e_v(H - Z_0 - Z_c) \\ -e_u \frac{W}{2} \end{pmatrix} = \begin{pmatrix} 0 & -v_1 \\ e_u & -u_1 \\ 0 & -v_2 \\ e_u & -u_2 \end{pmatrix} \begin{pmatrix} X_P \\ Y_P \end{pmatrix} + \underline{v} \tag{5}$$

C. Taking into account errors in the estimation process

Several kinds of errors are about to influence the estimation of the position using the above equation. The estimation explicitly takes into account the possible error on W, H, Z_0, u_1, v_1, u_2 and v_2 by including the covariance matrix of noise $\underline{v}$. This last matrix is calculated from the covariance matrix C_b of vector $\underline{B}$ grouping the contribution of all the parameters subject to error :

$$\underline{B} = \left(\Delta W, \Delta H, \Delta Z_0, \Delta u_1, \Delta v_1, \Delta u_2, \Delta v_2\right)^T \tag{6}$$

and from the Jacobian matrix D of system (5) :
So observation noise covariance matrix is

$$C_v = D C_B D^T$$

with

$$D = \begin{pmatrix} 0 & 0 & e_v & 0 & -Y_P & 0 & 0 \\ -\frac{e_u}{2} & 0 & 0 & -Y_P & 0 & 0 & 0 \\ 0 & e_v & -e_v & 0 & 0 & 0 & -Y_P \\ \frac{e_u}{2} & 0 & 0 & 0 & 0 & -Y_P & 0 \end{pmatrix}$$

$$C_{\mathbf{B}} = \begin{pmatrix} k^2\sigma_{b_H}^2 + \sigma_{b_W}^2 & k\sigma_{b_H}^2 & 0 & 0 & 0 & 0 & 0 \\ k\sigma_{b_H}^2 & \sigma_{b_H}^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{Z_0}^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{u_1}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{v_1}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{u_2}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_{v_2}^2 \end{pmatrix} \tag{7}$$

A Kalman filter has been used to realize the estimation of the pedestrian position. This estimator is well adapted to take into account error contributions as well as to allow prediction of the position.

This new estimator has proven to be highly reliable and will be the element for feeding adequately the tracker block (which is aimed at improving efficency of bounding box generation) in order to complete the eligible box's elaboration loop (see fig. 1). The association rule which has been initially designed exploits one box drop safety ring time-based, in order to maintain the label of the correct identifications through the frames of the acquired video.

V. Discussion

A new localization technique has been presented. This technique exploits scene localization of pedestrians by means of iterative image coordinates modeling. The reprojection onto the source omages shows a correct spatial positioning. The localization is not affected by target's movements.

The system has been tested in different situations. Currently, the result is not exploited by the preattentive phase, but results obtained by the localization phase promise to improve the reliability and efficiency of the preattentive stage.

A full tracking system that exploits the pedestrians localization function is currently under development.

Acknowledgments

The authors gratefully thank Dr. C. Adams and Dr. M. Del Rose from U. S. Army TACOM and Dr. S. Sampath from USARDSG-UK for their support in the research.

References

[1] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. von Seelen, "Walking Pedestrian Recognition," in *Procs. IEEE Intl. Conf. on Intelligent Transportation Systems'99*, (Tokyo, Japan), pp. 292–297, Oct. 1999.

[2] L. Zhao and C. Thorpe, "Stereo and neural network-based pedestrian detection," *IEEE Trans. on Intelligent Transportation Systems*, vol. 1, pp. 148–154, Sept. 2000.

[3] S. Kang, H. Byun, and S.-W. Lee, "Real-Time Pedestrian Detection Using Support Vector Machines," *Lecture Notes in Computer Science*, vol. 2388, p. 268, Feb. 2002.

[4] K. Fujimoto, H. Muro, N. Shimomura, T. Oki, Y. K. K. Maeda, and M. Hagino, "A Study on Pedestrian Detection Technology using Stereo Images," *JSAE Review*, vol. 23, pp. 383–385, Aug. 2002.

[5] D. M. Gavrila and J. Geibel, "Shape-Based Pedestrian Detection and Tracking," in *Procs. IEEE Intelligent Vehicles Symposium 2002*, (Paris, France), June 2002.

[6] D. M. Gavrila, "Sensor-based Pedestrian Protection," *IEEE Intelligent Systems*, vol. 16, pp. 77–81, Nov.–Dec. 2001.

[7] I. Haritaoglu, D. Harwood, and L. Davis, "W4: Who, When, Where, What: A Real Time System for Detecting and Tracking People," *Image and Vision Computing Journal*, vol. 17, Jan. 1999.

[8] D. Makris and T. Ellis, "Path Detection in Video Surveillance," *Image and Vision Computing Journal*, vol. 20, pp. 895–903, Oct. 2002.

[9] S. J. McKenna and S. Gong, "Non-intrusive Person Authentication for Access Control by Visual Tracking and Face Recognition," *Lecture Notes in Computer Science*, vol. 1206, pp. 177–184, Mar. 1997.

[10] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis and applications," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 781–796, Aug. 2000.

[11] C. Wöhler, J. K. Aulaf, T. Pörtner, and U. Franke, "A Time Delay Neural Network Algorithm for Real-time Pedestrian Detection," in *Procs. IEEE Intelligent Vehicles Symposium'98*, (Stuttgart, Germany), pp. 247–251, Oct. 1998.

[12] C. Wöhler, U. Kreßel, and J. K. Anlauf, "Pedestrian Recognition by Classification of Image Sequences – Global Approaches vs. Local Spatio-Temporal Processing," in *Procs. IEEE Intl. Conf. on Pattern Recognition*, (Barcelona, Spain), Sept. 2000.

[13] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. von Seelen, "Walking Pedestrian Recognition," *IEEE Trans. on Intelligent Transportation Systems*, vol. 1, pp. 155–163, Sept. 2000.

[14] V. Philomin, R. Duraiswami, and L. Davis, "Pedestrian Tracking from a Moving Vehicle," in *Procs. IEEE Intelligent Vehicles Symposium 2000*, (Detroit, USA), pp. 350–355, Oct. 2000.

[15] M. Bertozzi, A. Broggi, A. Fascioli, and P. Lombardi, "Vision-based Pedestrian Detection: will Ants Help?," in *Procs. IEEE Intelligent Vehicles Symposium 2002*, (Paris, France), June 2002.

[16] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context Free Attentional Operators: the Generalized Symmetry Transform," *Intl. Journal of Computer Vision, Special Issue on Qualitative Vision*, vol. 14, pp. 119–130, 1994.

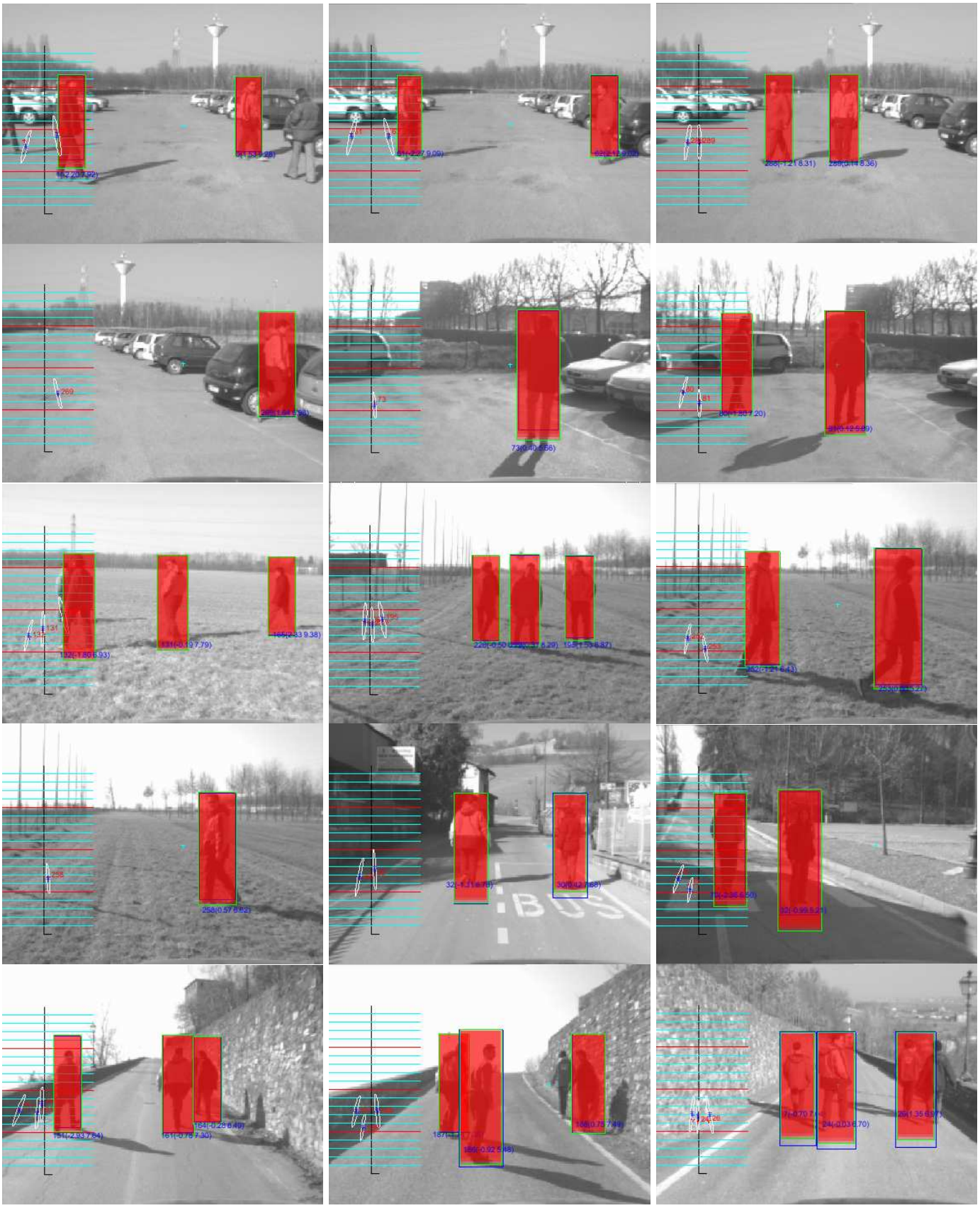

Fig. 6: Localization results: the localization area is superimposed on original images in red, the three numbers below these areas represent the localization ID and its coordinates in the world (meters). On the left side of the image a top view of the scene is sketched; each horizontal line represents 1 meter. For each pedestrian an error ellipsoid is given with its ID as well.