Real-time passenger counting in buses using dense stereovision

T Yahiaoui, L Khoudour, C Meurie

Real-time passenger counting in buses using dense stereovision

Tarek Yahiaoui
University Lille 1, Sciences and Technology
LIFL Laboratory
FOXMIIRE Team
IRCICA Building
Halley Avenue
Parc scientifique de la Haute-Borne
F-59650 Villeneuve d'Ascq, France
E-mail: tarek.yahiaoui@lifl.fr

Louahdi Khoudour
French National Institute for Transport and Safety Research
LEOST Laboratory
20 rue Elisee Reclus, BP 317
F-59666, Villeneuve d'Ascq Cedex, France

Cyril Meurie
University of Technology of Belfort-Montbeliard
Systems and Transportation Laboratory
ICAP Team
13 rue Ernest-Thierry Mieg
F-90010 Belfort Cedex, France

Abstract. *We are interested particularly in the estimation of passenger flows entering or exiting from buses. To achieve this measurement, we propose a counting system based on stereo vision. To extract three-dimensional information in a reliable way, we use a dense stereo-matching procedure in which the winner-takes-all technique minimizes a correlation score. This score is an improved version of the sum of absolute differences, including several similarity criteria determined on pixels or regions to be matched. After calculating disparity maps for each image, morphological operations and a binarization with multiple thresholds are used to localize the heads of people passing under the sensor. The markers describing the heads of the passengers getting on or off the bus are then tracked during the image sequence to reconstitute their trajectories. Finally, people are counted from these reconstituted trajectories. The technique suggested was validated by several realistic experiments. We showed that it is possible to obtain counting accuracy of 99% and 97% on two large realistic data sets of image sequences showing realistic scenarios.* © *2010 SPIE and IS&T.* [DOI: 10.1117/1.3455989]

1 Introduction

The considerable development of passengers traffic in public transportation has made it indispensable to set up specific methods of organization and management. For this reason, public transport companies are very much concerned with counting passengers,[1] which allows improved diagnosis of fraud, optimization of line management, traffic control and forecast, budgetary distribution between the different lines, and improvements in the quality of service. Therefore, developing a reliable passenger counting system becomes an important issue. Counting objects under controlled conditions, such as in manufacturing, is relatively easy, but counting people is much more difficult, especially under highly variable realistic environmental and operational conditions. Counting should be carried out with good accuracy, i.e., at least $\pm 3\%$ with a confidence rate of 95%. Accuracy and reliability should be consistently maintained throughout the counting process.

In France, several counting systems have been tested or are currently being tested in buses of the RATP, the Parisian transport operator. According to the results of these tests, the system must either be improved or replaced with a more accurate one. This is particularly necessary where fraud (people using buses without tickets) is concerned. The conclusion is that manual counting is carried out for one week every, on each bus line, in order to have an accurate evaluation of the traffic.

Nonetheless, technological progress has greatly improved systems of counting passengers. For example, the RATP has chosen a system with integrated infrared cells. Two types of cells, developed by ACOREL and ELINAP,

Paper 09128SSR received Jul. 17, 2009; revised manuscript received Dec. 7, 2009; accepted for publication Jan. 25, 2010; published online Dec. xx, xxxx.

1017-9909/2010/19(3)/1/0/$25.00

were initially tested by the RATP. These two solutions were not considered to provide sufficiently accurate counting. Thus, in 1996, a third type of cell, developed by BRIME, was considered to be sufficiently accurate and was installed in all the new vehicles.

Currently, RATP uses two types of automatic counting: ELINAP cells installed in 1500 vehicles (see http://www.acorel.com, for more details) and the BRIME systems installed in around 1000 vehicles (see http://www.brime-sud.fr, for more details). It is clear from this paragraph that RATP has been looking for automatic passenger counting systems for many years. The company has tested many of these without obtaining satisfactory results and now must carry out manual countings to readjust the automatic ones, which get less accurate over time. As far as we know, there are currently no systems in France that allow counting of passengers with an accuracy of $>95\%$ in buses. A study of the reliability of different systems of counting enables us to conclude that the two most reliable approaches:

1. The use of infrared directional sensors
2. Video sensing and image processing

Infrared directional sensors have a number of advantages, which explain their use in several systems of counting.[2] The major advantages are reduced size and cost, easy installation, and reliability. However, in crowded situations, their high sensitivity to noise, to variations in temperature, and to dust and smoke makes them less reliable in real-life situations. Moreover, they cannot distinguish between one passenger and a group of passengers, which is a huge drawback for counting in a bus. Thus, when counting passengers in a bus, a highly accurate system is necessary, particularly during rush hours. We believe that video-based systems are very promising for this task.

People counting using video is not a recent approach; we found in the literature many works dealing with this issue. The proposed techniques are various; however, based on their basic principle as a classification criterion, we distinguish the following classes:

1. Motion detection and analysis-based techniques: These can be described by a succession of two stages. The first one is to detect moving regions in the scene corresponding mostly to individuals. The second step uses the result of detection to rebuild over time, the trajectories of moving objects. The trajectory analysis is used to identify and count the people who crossed a virtual line or a predefined area.[3–5]
2. Edge analysis-based techniques: As their name suggests, these techniques exploit the extraction of edges for the detection. The objects of interest, in this case, correspond to a set of edges with a particular shape and organization. For example, a head corresponds to an edge with a circular shape.[6–8]
3. Model based techniques: These techniques attempt to find regions in the processed images that match predefined templates.[9,10] These models are either characteristics models or appearance models. The disadvantage of these approaches is either the need of a large learning database or a problem of model generalization.
4. Spatiotemporal techniques: These involve the selection of lines of interest in the acquired images and build on each line a space-time card by stacking lines in time. A second step is to use statistical models (templates) to derive the number of persons crossing the line and to analyze the discrepancies between the space-time maps in order to determine the direction.[11,12] These techniques have the advantage of being fast and simple to implement; however, works based on these techniques have not provided concrete solutions to interpret a significant number of cases. For example, the "blob" generated by a stationary person can be interpreted as that of several people.

Some researchers have been working in the field of counting people with monocular vision systems[13,14] and some with sets of video cameras scattered in the environment.[15,16] In the transport field, a system was developed by Mecoci *et al.*[17] to count passengers entering and exiting from buses. The authors claim that their system reaches a counting accuracy of 98%, but the evaluation presented in their paper was performed on a very reduced data set. Very few complete systems exploiting optical sensors and used in operation in transport context exist nowadays. Among these, we can mention the system developed by Albiol and Naranjo from Valencia University in Spain,[18] which provided interesting results. This system uses a single camera installed above the train doors of the RENFE railway network. The author announces a counting accuracy of 98% on realistic data sets corresponding to 149 train stops. The disadvantage of this system is that it mistakes an object and a large person, and the results are obtained using a correction factor. Given recent advances in computer vision and decreasing prices of hardware, the use of stereo vision is attractive. This approach is less sensitive to illumination changes and could also provide the necessary information to detect, model, and track objects or people. For all these reasons, we have chosen to develop a system based on dense stereo vision. However, we will see that stereo vision does not solve all the problems related to our application. In particular, the stereo matching could be very difficult for some cases.

This paper is organized as follows: In Section 2, we recall the basic aspects of stereo vision and show the interest of dense stereo vision for people counting. We also describe the hardware part of our system and present the overall structure of our image-processing chain. In Section 3, we present the similarity constraints enhancing the sum of absolute differences (SAD) score and compare the proposed stereo-matching technique with other methods on common images of the literature. Section 4 is devoted to the description of the other links of the processing chain: height map segmentation and feature tracking. In Section 5, we present the evaluation of our system on a laboratory data set, including various image sequences showing realistic scenarios, and on a real data set. Finally, a conclusion and a description of possible future work are provided in Section 6.

2 Stereovision for Counting Passengers

Stereo vision is a well-known method based on the analysis of several images (usually two) of the same object taken from different angles, along the optical axis of the camera

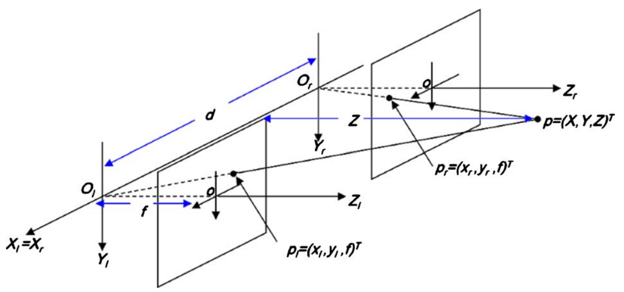

Fig. 1 Geometric modeling of binocular stereoscope.

(axial stereo vision), or by moving the acquisition system sideways (lateral stereo vision). Passive stereo vision operates a set of two (binocular vision) or three (trinocular vision) stereoscopic images.[19] It is static when observed objects do not move and dynamic where the objects can move.

In Section 2.1, we present the principle of the adopted binocular stereo vision. Then, we describe the hardware structure of the people-counting setup.

2.1 *Stereovision Vision Principles*

Figure 1 shows a typical stereo-vision setup, in which optical axes of the two cameras are parallel. The distance d between these optical axes is called the baseline of the stereo-vision setup. It is generally assumed that the two cameras have exactly the same focal distance f. A region of the scene exists in which points are visible by both cameras. In the image-formation process, a point P of this region is projected onto a pixel P_l of the image sensor of the left camera and onto a pixel P_r of the image sensor of the right camera. Pixels P_l and P_r are called homologous because they correspond to the same point of the scene. The disparity is defined as the difference between horizontal positions of homologous pixels; the further the point P is from the cameras, the smaller the disparity is. Stereo-vision techniques aim at recovering various information about the real scene using only the visual data contained in the two images. This problem is not trivial since the pairs of homologous pixels are not known *a priori*.

Usually, stereo-vision techniques include two parts: stereo matching and 3-D reconstruction. For passenger counting in buses, because the sensor is very close to persons passing under it, it is difficult to extract particular points (such as curves) and segments, and to match them. We have tested some well-known sparse stereo-vision algorithms on our data set,[20–22] without success for features extraction. With a dense stereo approach, we will show later that it is possible to reconstruct a height map, in which the heads of people can be easily located.

2.2 *Our People Counting System*

The global system is composed of an acquisition part and a processing part. The acquisition device is an industrial stereoscopic sensor called bumblebee (manufactured by the PointGrey Company), fixed vertically above the entrance of the bus at a height of 235 cm with a baseline of 12 cm. The processing chain, which counts people passing under the system using the images acquired by the hardware setup, is composed of the following links:

1. A stereo-matching block that computes the disparity map for each pair of images. This map is then transformed into a height map for further processing.
2. A segmentation block that identifies, in the height map, heads of people by detecting round shapes with a constant height value.
3. Tracking and counting modules that reconstruct the trajectories of people's heads using the round shapes marked in successive stereo pairs. A person is counted by this module when the trajectory of his/her head enters or leaves the stereo field of view.

The key point of this processing chain is the computation of precise and accurate height maps. The proposed dense stereo-matching approach is described in Section 3. The other steps of the processing chain (i.e., segmentation and marker tracking for trajectory reconstruction) will be described later.

3 Improved Stereo Matching

3.1 *Principles of SAD Matching Cost*

The dissimilarity measure, also called correlation, is one of the most widely used techniques for determining all the homologous pixels. It consists of defining a neighborhood, around each pixel of the right image, and measuring the ressemblance between it and the same neighborhoods surrounding pixels of the left image. We calculate for each pixel of the left image a dissimilarity curve as a function of the shift that defines the minimum and maximum disparities allowed by the imaging system. In the case of the SAD matching cost [winner-takes-all (WTA) algorithm],[23,24] the dissimilarity measurement corresponds to the absolute difference defined by Eq. (1). Thus, the shift corresponding to the minimum value of the dissimilarity curve marks the pixel supposed to be the homologous one of the pixel of the left image that we try to match,

$$C_{\mathrm{SAD}}(x,y,s) = \sum_{ij} \left| G(x+i+s,y+j) - D(x+i,y+j) \right|. \qquad (1)$$

where $G(x,y)$ is the gray level of the pixel (x,y) we want to match and that belongs to the left image, $D(x,y)$ is the gray level of the pixel (x,y) in the right image, s is the shift between the two pixels (left and right), and d is the disparity that corresponds to the shift-minimizing C_{SAD} criterion defined in Eq. (1).

The advantage of the SAD matching cost (WTA algorithm) described above is that it is simple to implement, robust and fast enough to operate in real time.[25] However, some matching errors are caused by this approach, which leads to an incorrect disparity value on some given pixels. In addition, one of the major drawbacks of this method is to systematically yield a matching result even if the area of the scene is partially or totally occluded, in which case these results are false. Thus, in order to reduce the number of matching errors, we propose an approach, based on the SAD matching cost (WTA algorithm), in which we impose constraints for the selection and better matching of the neighborhoods.[26] This improves the matching, taking into account various types of areas: hidden, not hidden, and under the influence of illumination changes.

3.2 *Improvements Brought to the SAD Matching Cost (WTA Algorithm)*

Four similarity constraints are introduced to improve the matching process with the WTA algorithm.

3.2.1 *Similarity of the gray levels of pixels to be matched*

The first similarity criterion between two homologous pixels is the similarity of their gray levels. When using square or symmetric rectangular neighborhoods, we consider the pixel to match as the center of the first calculation neighborhood, called fixed, and the candidate pixel as the center of the second calculation neighborhood, called sliding. The aim of this constraint is to increase the matching accuracy by promoting the matching of the most similar pixels. This is achieved by promoting a minimum compared to others in the case of multiple minima of the dissimilarity curve (for example, in the case of repetitive textures). We call α the coefficient assigned to this similarity criterion. This coefficient can take only two values, depending on whether the constraint is introduced or not. We look for the pixel that minimizes the dissimilarity criterion of Eq. (2). Thus, for a shift satisfying the constraint, the introduction of the coefficient α will further minimize the value of dissimilarity. We propose a simple multiplication of the coefficient α and the dissimilarity term of Eq. (2). Let us call this expression C_1. In order to make the overall term lower when the constraint is introduced, it is necessary that the particular value that α takes when the constraint is introduced be <1.

$$C_1(x,y,s) = \alpha \times \sum_{ij} \left| G(x+i+s,y+j) - D(x+i,y+j) \right|, \quad (2)$$

where $\alpha=1$ if the constraint is not verified and $\alpha=\alpha_0$ knowing that $0<\alpha_0<1$, if the constraint is introduced. We consider that the constraint is introduced if the difference between the gray levels does not exceed a given threshold, fixed experimentally.

3.2.2 *Stereo matching of pixels belonging to identified edges*

We also use an additional similarity criterion to deal with the matching of edge pixels. These pixels have a higher probability to correspond to regions of hidden areas or near-hidden (occluded) regions. Usually, in stereo vision, we can reasonably assume that if a pixel corresponds to an edge, so does the homologous pixel. On the basis of this assumption, we can introduce this constraint to try to improve the matching of pixels corresponding to these edges. Edge pixels are extracted using a classical Laplacian-based technique.[27] Because of the difficult application environment (occlusion, high illumination variation), good detection is hard to achieve. However, even though it is not perfect, we use this information. Therefore, there is no need to develop a complex approach to obtain it. As with the previous constraint, we have associated a weighting factor called β to this similarity criterion. Let us call the expression linked to this constraint C_2

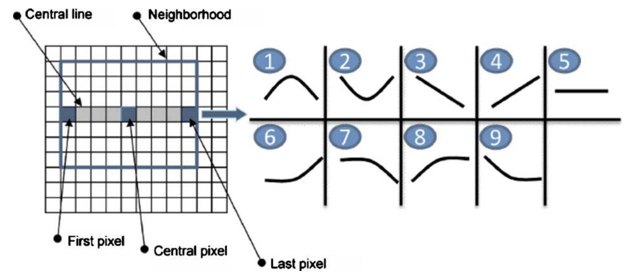

Fig. 2 Profiles for the gray levels of the pixels belonging to the central lines of the calculation neighborhoods.

$$C_2(x,y,s) = \beta \times \sum_{ij} \left| G(x+i+s,y+j) - D(x+i,y+j) \right|, \quad (3)$$

where $\beta=1$ if the constraint is not introduced and $\beta=\beta_0$ knowing that $0<\beta_0<1$, if the constraint is introduced.

3.2.3 *Similarity of simplified gray-level profiles of the pixels corresponding to the centerlines of calculation neighborhoods*

We define an additional similarity criterion in analyzing simplified gray-level profiles of the pixels of the center lines of the two calculation neighborhoods. Figure 2 provides the main simplified gray-level profiles for a given window size. The gray level profiles of the center lines of the two calculation neighborhoods are analyzed and compared. If the two gray-level profiles correspond to homologous pixels, the two-gray-level curves should have the same profile.

We associate to this new constraint the weighting factor γ. Let us call the expression linked to this new constraint C_3,

$$C_3(x,y,s) = \gamma \times \sum_{ij} \left| G(x+i+s,y+j) - D(x+i,y+j) \right|, \quad (4)$$

where $\gamma=1$ if the constraint is not introduced and $\gamma=\gamma_0$ knowing that $0<\gamma_0<1$, if the constraint is introduced.

3.2.4 *Use of motion*

The motion-detection approach is based on the substraction of a background image. The motion detection is carried out for both images. Before matching, we classify the pixels of the left and right images into two classes, based on whether or not the pixels belong to regions affected by motion. The basic idea is to introduce, as with the previous similarity constraints, a coefficient called μ in the dissimilarity criterion (called C_4). This coefficient will favor homologous pixels belonging to the same class of regions: moving or static. This also drastically lowers the computation time by matching only pixels belonging to moving areas,

$$C_4(x,y,s) = \mu \times \sum_{ij} \left| G(x+i+s,y+j) - D(x+i,y+j) \right|, \quad (5)$$

where $\mu=1$ if the constraint is not introduced and $\mu=\mu_0$ knowing that $0<\mu_0<1$, if the constraint is introduced.

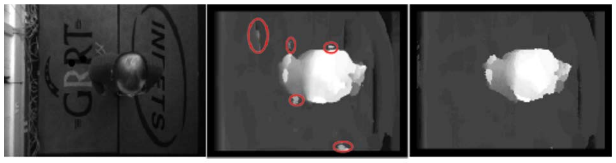

Fig. 3 Example of disparity maps calculated on a pair of images: (a) Left image, (b) SAD, and (c) our method.

3.2.5 *Associations of constraints*

Thus far, we have proposed four similarity constraints to improve the accuracy of pixel matching. Knowing that each of these constraints is of a different nature, it becomes interesting to combine these various similarity criteria to increase the robustness of the matching process and analyze their respective values. In other words, we simultaneously do the following:

1. Compare the similarity or dissimilarity of neighborhoods corresponding to the pixel to match and the candidate pixel
2. Check if their gray levels are similar
3. Test if they belong to edges
4. Verify whether the gray-level profiles of central lines of calculation neighborhoods are similar
5. And, finally, test if they both belong to a region affected by motion

We can find in the literature diverse techniques allowing the association of several criteria in order to optimize a global one. The most used optimization criteria are based on genetic algorithms,[28] fuzzy logic,[29] analysis of variance,[30] decision trees,[31] and derivative approaches.[32] The optimization technique choice should meet a compromise between the complexity of the problem to solve and the optimization result.

In our case, we consider that the similarity criteria are of a different nature and are more or less independent. Thus, we chose to use an additive model for the calculation of dissimilarity, which corresponds to summing the dissimilarity of four criteria,

$$C(x,y,s) = C_1(x,y,s) + C_2(x,y,s) + C_3(x,y,s) + C_4(x,y,s), \tag{6}$$

where C_1, C_2, C_3, and C_4 match dissimilarity in the order they were presented. The global formulation becomes

$$C(x,y,s) = (\alpha+\beta+\gamma+\mu) \times \sum_{ij} |G(x+i+s,y+j) - D(x+i,y+j)|. \tag{7}$$

Figure 3 provides two disparity maps calculated with the SAD alone and with the four constraints together, on a pair of stereoscopic images. We not that for SAD some matching errors appear (marked with ellipses). This visually shows the improvement brought by the introduction of constraints in SAD model.

To test the relevance of our algorithm, we compared our approach to classical approaches having the same complexity and calculation time as ours. We retained methods using the following statistical distances: SAD, zero mean SAD, sum of squared differences (SSD), and zero mean SSD. The algorithms with which we conduct a comparison are those proposed by Scharstein and Szeliski.[33] In the framework of this paper, we only provide results on the evaluations of the first three constraints (C_1, C_2, and C_3) because we only have single images with ground truth and thus cannot compute motion. Therefore, the C_4 constraint, which requires motion detection, is not used in this comparison. The first stereoscopic images of the test are a couple of synthetic images (Corridor of Lena in Fig. 4). The second stereoscopic pair is relatively difficult to match because of the complex and repetitive textures (Cones in Fig. 4). The third stereoscopic pair of images is a view of a natural scene. The main difficulties of matching pixels of this pair of images is a highly textured background and many occlusions (Tsukuba in Fig. 4). In Fig. 4, for each case, we show left and right images and the disparity map representing the ground truth.

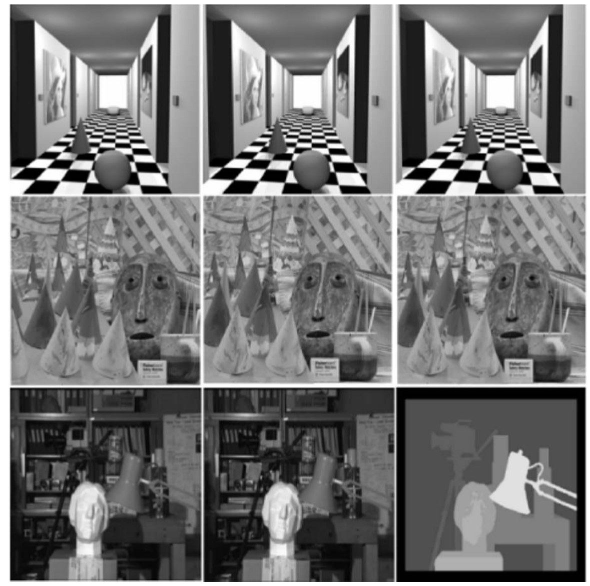

Fig. 4 Pair of stereoscopic images for comparison: (a) Corridor of Lena, (b) cones, and (c) Tsukuba.

Our algorithm is compared to SAD matching cost (WTA algorithm) and its family following two criteria: with the ground truth, we calculate the number of pixels correctly matched to the total number of candidate pixels. This is achieved separately for occluded and nonoccluded pixels. For each pair of images tested, the best values of the parameters α_0=0.85, β_0=0.85, γ_0=0.90, and μ_0=0.80 with a neighborhood of 15×15 pixels. The coefficients and neighborhood values corresponding to those minimize the matching-error rate curves. The overall results are as follows:

1. Each of the constraints taken independently from the others reduces the matching error rate of mapping.
2. By combining the three constraints, we obtain the best results.
3. By varying the size of the calculation neighborhood from 3×3 pixels to 21×21 pixels, the matching er-

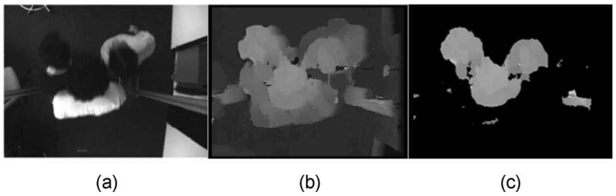

Fig. 5 Artifacts elimination by morphological filtering: (a) Left image, (b) disparity map, and (c) result of smoothing.

ror rate decreases to reach a minimum corresponding to an average calculation neighborhood size (often 15×15 pixels), and then it increases. The effect of the three constraints together on the real Cones and Tsukuba images (gain of 3%) are the most important, especially on occluded pixels.

4 Segmentation and Tracking

In Section 3.2, we described an improved stereo-matching method that allows the computation of precise and noise-free height maps. These maps are segmented in order to detect heads of people, and the marked areas are tracked across the image sequence.

In Fig. 5, we can see the processing carried out and the results obtained: for a given disparity map in Fig. 5(b), a threshold is first applied to retain only the parts of the image close to the camera; the result is displayed in Figs. 5(c) and 6(a). Then, a binarization and size-based artifact removal yields the binary image in Fig. 5(b). One more processing step is necessary to highlight the heads of people. For this, we use binary mathematical morphology. Three opening operations are applied to the binary images with a circular structuring element. As with every morphological filtering, the size of the structuring element is very important. The result is shown in Fig. 6(c). We can see in Fig. 6(a) that the majority of the artifacts have disappeared. The result is satisfactory because we get three different kernels corresponding exactly to the heads of the persons if we compare to the original images.

For a given stereo configuration, we can define a statistical average size of a head on the image as a function of the distance that separates the human head from the cameras. This means that we cannot use the same structuring element for segmenting heads of people having different heights. To deal with this problem, we define several height intervals corresponding to different height classes. For each class, we use a specific structuring element having a size equivalent to the average size of a head, based on the height and, therefore, on the distance from the camera. Given the variability of people's heights, defining the number of height classes is not easy. This number has a strong influence on the quality of the result; thus, it must be chosen carefully. It must be large enough to represent the majority of people's height classes and not too large to avoid increasing the processing time. Experimentally, we found that four classes are a good compromise.

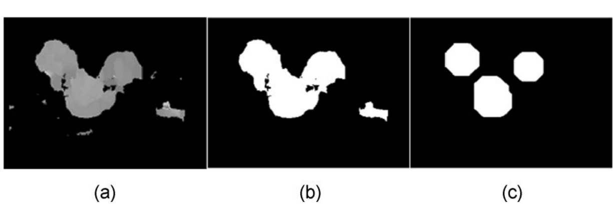

Fig. 6 Use of binary mathematical morphology for the disparity map segmentation: (a) Result of smoothing of the previous step, (b) binary image, and (c) kernels results.

These classes are used for thresholding the disparity map, and in the same way as shown in Fig. 6, morphological tools are then applied to each thresholding result to segment the heads of people. For a given class, the size of the kernels resulting from this segmentation step leads to differentiate objects larger than the average head size of the class. Then, the differentiation between large objects and head is carried out by the tracking procedure.

The tracking of the kernels for the final counting is performed using a Kalman filter.[34] Each kernel resulting from the segmentation of the disparity maps is represented by a vector of the following seven components:

1. Number of pixels
2. Width of the kernel in pixels
3. Length of the kernel in pixels
4. Average height calculated from the heights of each pixel
5. Average gray level
6. Abscissa in the image
7. Ordinate in the image

The main aim of the tracking algorithm in this case is to track the kernels in the processing zone (called also counting zone) and to analyze the behavior of the kernels (which are, in fact, the heads of the persons passing under the sensor) in the counting zone. The first step of the tracking procedure is the multitarget Kalman filter, which provides prediction of kernels positions. We assume that each target is represented by a vector $\boldsymbol{X}$ of two components (x,y), where x and y are the horizontal and vertical coordinates of kernels in the image. The prediction is made based on two assumptions: the speed of objects is constant and the measures are affected by white noise. The second step corresponds to the calculation of a probability mapping. In this step, the estimation of the probabilities requires the prediction from Kalman filter, corresponding to horizontal and vertical coordinates of the targets, and the five others kernel parameters used without prediction. These probability measures are also weighted by tracking hypotheses (merging, splitting, appearance, disappearance, …). A similar tracking methodology is described in Ref. 34. We introduce, then, the notion of trajectory. A valid trajectory corresponds to somebody entering and exiting from the counting zone. The counting zone has an upper and lower line; the interior is called the tracking zone.

The valid trajectories corresponding to an entry in the counting zone are the following [Fig. 7(a)]:

1. Appearance of a person at the upper line of the counting zone and disappearance in the tracking zone (the person has entered and stays in the tracking zone: they are taken into account)
2. Appearance at the upper line of the counting zone and disappearance at the lower line of the counting

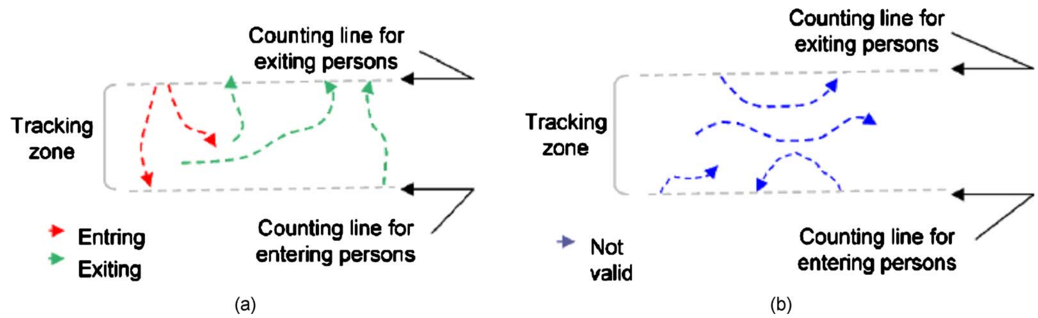

Fig. 7 Examples of (a) valid and (b) nonvalid trajectories.

zone (the person entered and crossed the counting zone: they are counted).

The nonvalid trajectories are linked to the following situations [Fig. 7(b)]:

1. Appearance at the upper line of the counting zone and disappearance at the same line (entry followed by an immediate exit)
2. Appearance at lower line and disappearance at the same line
3. Appearance and disappearance in the counting zone (wandering under the sensor without intention)
4. Appearance at lower line and disappearance in the tracking zone

5 Evaluation of the Counting System

The overall evaluation of the system is carried out following two directions. First of all, we are interested in the performance of the system by comparing globally the results of the counting system to ground truth determined by several experts. It is a quantitative evaluation. Then, because the counting is based on the notion of valid trajectories, a qualitative evaluation is also carried out in order to analyze the ability of the system to manage difficult situations.

5.1 *Data Sets Used for the Evaluation*

First of all, let us mention that the counting system was entirely evaluated on real data sets. The data sets on which the system was evaluated come from two different data bases. In the framework of this paper, the data used for the evaluation includes 30 laboratory scenarios and 96 scenarios coming from a bus.

Laboratory data respecting specific scenarios was provided by the RATP, and 30 scenarios were simulated in our laboratory. They reflect mainly situations where people are exiting from a bus. The scenarios represent very diverse situations: high-density groups of people moving in opposite directions; people of different sizes, carrying bags, suitcases, or big objects; and people with strollers. One should note here that the position of the sensor and the choice of the focal length of the lens were chosen to reproduce exactly the geometrical aspects of the bus. The first 15 scenarios were simulated with ambient illumination (artificial light and daylight coming from the windows), whereas the must 15 were played with closed windows and artificial light shut off.

Real data coming from a bus during the exploitation period lasted for one day, on a very crowded line. The collected data represent various situations: crowd, strollers, luggage, children, and people with hats; 150 scenarios of these typical situations were collected. The processing time

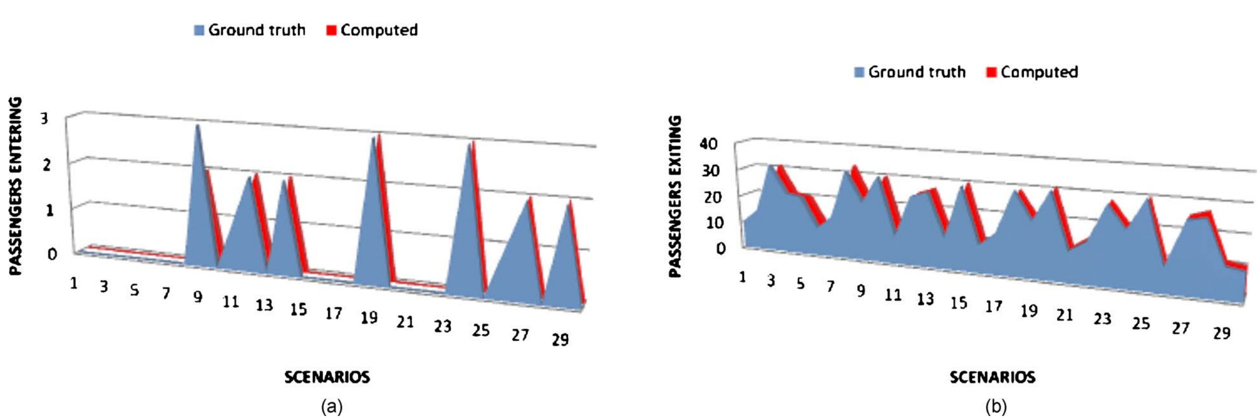

Fig. 8 Counting results for 30 scenarios in laboratory (from top to bottom): (a) entering and (b) exiting by the same door.

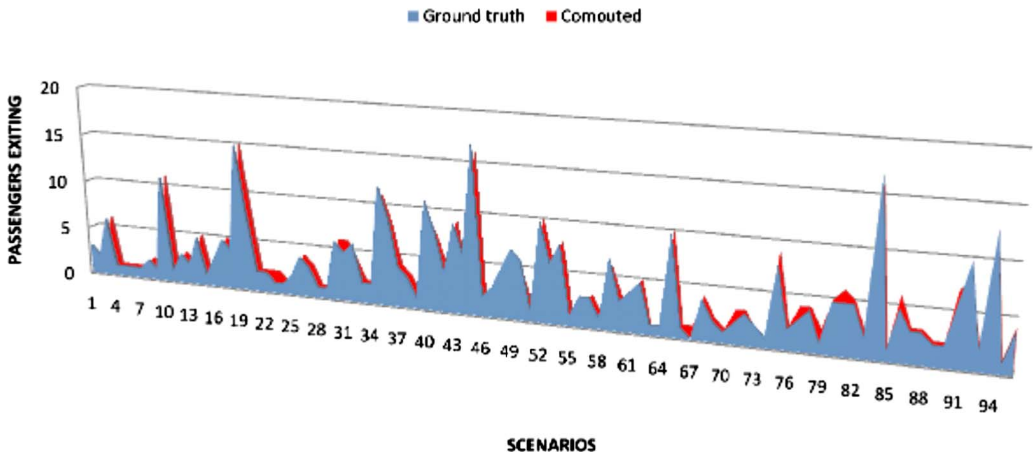

Fig. 9 Counting results for 96 scenarios in a bus.

is 30 fps if we consider images whose resolution is 160 $\times$120 pixels on a pentium IV 2 GHz. This is compatible with our application.

5.2 *Quantitative Evaluation*

The counting results presented in Fig. 8 indicate the number of people entering or exiting for each sequence in the laboratory. In Fig. 8, we can see the ground-truth counting results versus the counting results computed by our algorithm. One can note that whatever the difficulty of the scenario is, the difference between the reference and calculated countings is very low. Indeed, these differences are in the interval $[-1;+1]$. This is an encouraging result showing the robustness of our algorithm, which is able to cope with diverse situations. There are fewer people entering because the data set corresponds mainly to people exiting by the back door, and there are counting errors because people are entering and exiting at the same time by the same door.

In order to determine the accuracy of our counting system, globally—that is to say considering all the entering and exiting scenarios together—we have defined an error rate that is calculated with Eq. (8). In this equation, we consider the real counting (the ground truth obtained with three different experts) as the basis of comparison and determine the difference between the counting with the algorithm. Thus, the error rate is $\sim$1%,

$$\text{Error}_{\text{counting}} = 100\frac{(\text{Real}_{\text{counting}} - \text{Automatic}_{\text{counting}})}{\text{Real}_{\text{counting}}}. \tag{8}$$

The same error rate is obtained with any laboratory scenario, under any illumination type. This is also encouraging. For the bus data sets, the results are shown in Fig. 9. We can note in Fig. 9 that the ground-truth results are very close to the results after computation with our algorithm. Even though the scenarios are much more difficult to deal with in the bus, the overall counting error is only 3%. When analyzing more closely the counting results, we observe that when our system differs from the reference counting, it systematically underestimates the number of people. Several reasons could explain this fact: the difficulty to detect short people. The fixed size of the structuring element in the segmentation of the disparity maps could also be another reason. Finally, the merging of two trajectories, corresponding to two different people could also be an additional reason. Additional explanations could also be found with a more intensive evaluation.

5.3 *Qualitative Evaluation of the Counting System*

After the quantitative evaluation of the system, it is interesting to carry out qualitative evaluation of the algorithm on typical image sequences. The main aim of this section is to show the behavior of the counting system on different trajectories of people passing under the sensor. The objective is also to verify the ability of the system to detect specific people, to track them, and finally to count them. To achieve this goal, we have selected three typical sequences: two from laboratory data sets and one from a bus in normal operation. For each sequence, we present the following conclusions.

Sequence 1 represents a crowd exiting from the counting zone while at the same time, several other people are entering one behind the other (Fig. 10). The main interest of this sequence is to show the ability of the system to analyze the trajectories of people having the same characteristics in terms of size and appearance. We have marked people under analysis, with color ellipses: red for people exiting and green for people entering.

Sequence 2 illustrates two people walking very close to each other. One person puts his arm on the shoulders of the other. This situation is illustrated in Fig. 11 in four frames. As for the previous sequence, the heads are marked with red ellipses. The two persons are exiting from the counting zone.

Sequence 3, which is acquired in the bus, represents a crowd getting off the bus. Among this crowd are several children, and several other people are standing at the entrance without leaving the bus (typical situation in buses).

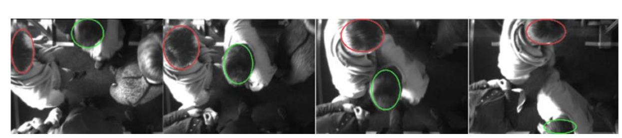

Fig. 10 Images taken from sequence 1: Evolution in time.

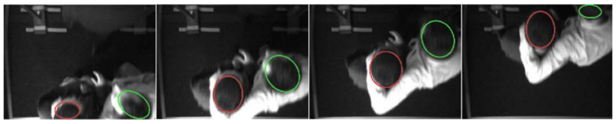

Fig. 11 Images taken from sequence 2: Evolution in time.

The main interest of the sequence is to test the ability of the system to detect a young child, a stationary person, and a person wearing a hat. Figure 12 illustrates this situation. The green ellipse indicates the stationary person; the red one, the child exiting from the bus; and the blue one, the man with the hat who is also exiting from the bus.

5.3.1 *Tracking results*

The tracking results are illustrated in Figs. 13–15. The colors used for drawing the trajectories are those used in Figs. 10–12.

In Fig. 13, which corresponds to sequence 1, we have represented the trajectory of the person entering in continuous line and the trajectory of the person exiting in dashed line. The abscissa and ordinate in the graph represent the spatial position, of the centers of gravity of the heads of the passengers, in the counting area, detected during the segmentation phase. Every kernel is calculated at 30 fps, but the center of gravity is plotted only every five frames for visual convenience. We note that, in spite of the high proximity of the two people, the respective trajectories are perfectly identified: one entering and the other exiting. We can also note that the trajectory of the person entering is more rectilinear than that of the exiting person because the latter has diverted his trajectory in order to avoid a collision.

In Fig. 14, we can note that the system has perfectly dealt with the typical situation where two people are crossing the counting zone very closely. We can clearly distinguish two parallel trajectories describing their passage.

In Fig. 15, we can easily note the trajectory (dashed line) of the kid who has rapidly gotten off the bus. The continuous line corresponds to the man with the hat. For this person, in spite of the lack of contrast between his clothes and the background, the system has detected the trajectory properly. The third trajectory is typical of people standing at the exit of the bus but moving a little, from time to time, to let the other passengers get off the bus. That is why the position of the center of gravity of the head moves slightly. In Fig. 15, because the child and the man with the hat are getting off the bus, one behind the other, the corresponding trajectories are almost aligned.

5.4 *Real-Time Constraints*

The first version of the algorithm was implemented on a PC Pentium IV 2 GHz and processed images of size 640 $\times$480 pixels. But, with this size, the algorithm was only

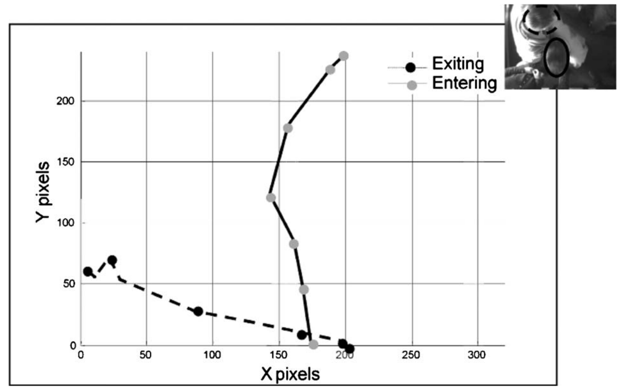

Fig. 13 Trajectories of people marked in sequence 1.

able to process up to 2 fps, and it was impossible to count people moving very quickly. The real-time constraints for this system are the following: Every person must be counted, regardless of their speed of movement. A processing time of 2 fps cannot be considered real time.

Therefore, in order to speed up the processing time, we tried to reduce the size of the images while striving to maintain the accuracy. Then, we tested two images sizes: 320×240 and 160×120 pixels. We have concluded that the best compromise, in terms of accuracy and processing time, was achieved by an image size of 160×120 pixels. In this case, the accuracy is maintained and the processing speed is 30 fps, which is compatible with a real-time implementation. The accuracy is not affected when we divide the resolution by four moving from 640×480 to 160 $\times$120 pixels, which demonstrates the robustness of the algorithm proposed.

6 Conclusion

In this paper, we have presented a counting system and its evaluation on life-situation data sets. The comparison between ground-truth values and the ones calculated with our algorithm leads to a counting accuracy that is around 99% for laboratory and 97% for bus data sets. These values are obtained on 30 scenarios coming from the laboratory and 96 coming from a bus during the exploitation period and representing a total of ~1400 people. This counting accuracy needs to be confirmed with a more intensive evaluation, mainly on the scenarios coming from the bus. We have also conducted a qualitative evaluation in order to test the ability of our algorithm to detect and track persons and their trajectories in a few very difficult situations. We have tested the robustness of the algorithm to deal with very hard cases: very crowded situations where there are people walking in two directions under the sensor.

The results obtained in these cases are very satisfactory and encourage conducting us to continue working in this

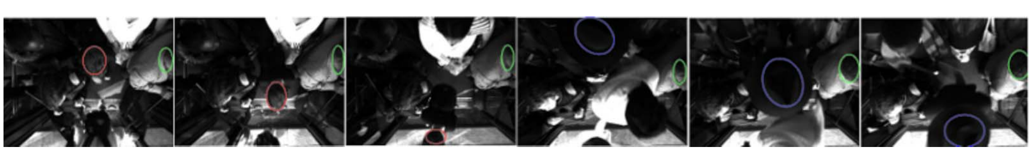

Fig. 12 Images taken from sequence 3: Evolution in time.

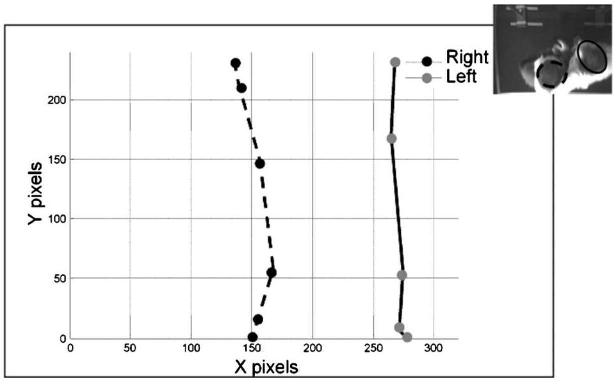

Fig. 14 Trajectories of people marked in sequence 2.

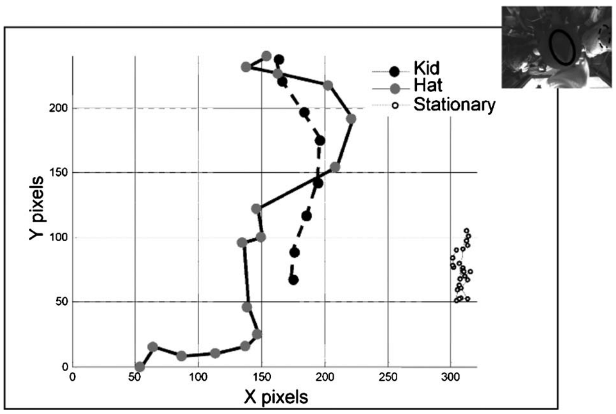

Fig. 15 Trajectories of people marked in sequence 3.

direction. That is why numerous perspectives are planned in the near future. We plan, for instance, to separate the data to assess the results in crowded situations versus non-crowded ones. Because we wanted a real-time counting system, from the beginning, the use of color images was avoided because of the extra processing time they imply. However, the use of color would provide improvements in the choice of homologous pixels for the stereo-matching process because we have more information for neighborhood comparison. Finally, color information could be used to perform pixel clustering of the stereoscopic images in a number of classes which could be then exploited. For instance, we could imagine adding additional constraints depending on the classification results.

Acknowledgments

We thank the Paris transport operator (RATP: Regie Autonome des Transports Parisiens) who funded this research carried out by means of a Ph.D. thesis. This collaboration between RATP, INRETS, and USTL (LAGIS laboratory, University Lille 1, Sciences and Technology) was really fruitful. This counting system was patented by the French organization CNISF: National Council of Engineers and Scientists of France under Grant No. 0953188.

References

1. D. Beymer, "Person counting using stereo," in *Workshop on Human Motion*, pp. 127–133, IEEE Computer Society, Washington, DC (2000).
2. A. Pincot, "Fiabilisation de la chaine de comptage voyageurs," RATP, Tech. Rep. RATP-MRB (Mar. 2002).
3. X. Liu, P. Tu, J. Rittscher, A. Perera, and N. Krahnstoever, "Detecting and counting people in surveillance applications," *Proc. IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 306–311, IEEE Computer Society, Washington, DC (2005).
4. E. Zhang and F. Chen, "A fast and robust people counting method in video surveillance," in *CIS '07: Proceedings of 2007 Int. Conf. on Computational Intelligence and Security*, Washington, DC, pp. 339–343, IEEE Computer Society, Washington, DC (2007).
5. X.-W. Xu, Z.-Y. Wang, Y.-H. Liang, and Y.-Q. Zhang, "A rapid method for passing people counting in monocular video sequences," in *Proc. of 6th Int. Conf. on Machine Learning and Cybernetics*, Hong Kong, pp. 1657–1662, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, WI (2007).
6. M. Bozzoli and L. Cinque, "A statistical method for people counting in crowded environments," *Proc. 14th International Conference on Image Analysis and Processing (ICIAP 2007)*, pp. 506–511, IEEE Computer Society, Washington, DC (2007).
7. A. Gardel, I. Bravo, P. Jimenez, J. Lazaro, and A. Torquemada, "Real time head detection for embedded vision modules," in *Proc. of IEEE Int. Symp. on Intelligent Signal Processing (WISP 2007)*, pp. 1–6, IEEE, Washington, DC (2007).
8. S. Yu, X. Chen, W. Sun, and D. Xie, "A robust method for detecting and counting people," in *Proc. of Int. Conf. on Audio, Language and Image Processing (ICALIP 2008)*, pp. 1545–1549, IEEE, Washington, DC (2008).
9. O. Sidla, Y. Lypetskyy, N. Brandle, and S. Seer, "Pedestrian detection and tracking for counting applications in crowded situations," in *AVSS '06: Proceedings of the IEEE Int. Conf. on Video and Signal Based Surveillance*, Washington, DC, IEEE Computer Society, pp. 70–75 (2006).
10. G. García-Bunster and M. Torres-Torriti, "Effective pedestrian detection and counting at bus stops," in *Proc. of Robotic Symp., IEEE Latin American*, pp. 158–163, IEEE, Washington, DC (2008).
11. J. Barandiaran, B. Murguia, and F. Boto, "Real-time people counting using multiple lines," in *Proc. 9th Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '08)*, Washington, DC, IEEE Computer Society, pp. 159–162, IEEE Computer Society, IEEE Computer Society, Washington, DC (2008).
12. Y. Jeon and P. Rybski, "Analysis of a spatio-temporal clustering algorithm for counting people in a meeting," Robotics Institute, Pittsburgh, Tech. Rep. No. CMU-RI-TR-06-04 (Jan. 2006).
13. G.-P. Adriano, S.-I.-V. Mendoza, F.-N.-J. Montinola, and P.-C. Naval, "Apec: Automated people counting from video," presented at PCSC Conf. Security and Networking (2005).
14. V. Rabaud and S. Belongie, "Counting crowded moving objects," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 705–711 (2006).
15. A.-O. Ercan, A. E. Gamal, and L.-J. Guibas, "Object tracking in the presence of occlusions via a camera network," in *Proc. of 6th Int. Conf. on Information Processing in Sensor Networks*, pp. 509–518, ACM, New York (2007).
16. S. Fleck, C. Vollrath, F. Walter, and W. Straber, "An integrated visualization of a smart camera based distributed surveillance system," in *Proc. of 3rd Conf. on IASTED Int. Conf.: Advances in Computer Science and Technology*, pp. 234–242, ACTA Press, Anaheim, CA (2007).
17. A. Mecoci, F. Bartolini, and V. Cappellini, "Image sequence analysis for counting in real time people getting in and out of a bus," *Revue Signal Process.* **35**, pp. 105–116 (1994).
18. A. Albiol, V. Naranjo, and I. Mora, "Real-time high density people counter using morphological tools," *IEEE Trans. Intell. Transp. Syst.* **3**, 204–217 (2001).
19. S. Bahrodi, L. Iocchi, G. Leone, D. Nardi, and L. Scozafava, "Real-time people localization and tracking through fixed stereo vision," *Rev. Appl. Intell.* **26**(2), 83–97 (2007).
20. Y. Zhang and C. Kambhamatteu, "Stereo matching with segmentation-based cooperation," in *Proc. of 7th European Conf. on Computer Vision*, Vol. 2, pp. 556–571, In Lecture Notes in Computer Science, vol. 2351/2002, pp. 521-522, Springer, Heidelberg (2002).
21. Y. Ruichek and J.-G. Postaire, "A new neural real-time implementation for obstacle detection using linear stereo vision," *Real-Time Imaging J.* **5**, 141–153 (1999).
22. Y. Ruichek, H. Issa, and J.-G. Postaire, "Genetic approach for obstacle detection using linear stereo vision," in *Proc. of IEEE Intelligent Vehicles Symp.*, pp. 261–266 (2000).
23. S. Haris, W. Vandermark, and D.-M. Cavrila, "A comparative study of fast dense stereo vision algorithms," in *Proc. of IEEE Intelligent Vehicles Symp.*, pp. 319–324 (2004).
24. S. Wong, S. Vasiliadis, and S. Cotofana, "A sum of absolute differences implementation in FPGA hardware," in *Proc. of 28th EUROMICRO Conf.*, pp. 183–188, IEEE, Washington, DC (2002).
25. T. Yahiaoui, F. Cabestaing, L. Khoudour, and P.-H. Leny, "Le comptage de passagers entrant et sortant d'un autobus par stéréovision dense," presented at Int. Workshop: Logistique & Transport

(2006).
26. T. Yahiaoui, "Une approche de stéréovision dense intégrant des contraintes de similarité. application au comptage de passagers entrant et sortant d'un autobus," Ph.D. dissertation, University of Lille (2007).
27. H. Maitre, Détection de contour dans les images, ⟨http://www.tsi.enst.fr/~bloch/TDI/poly_contours.pdf ⟨⟩accessed July 2, 2010).
28. D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Kluwer, Dordrecht (1989).
29. L.-A. Zadeh, G.-J. Klir, and B. Yuan, *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems*, World Scientific, 1996.
30. R.-W. Pike, Optimization for engineering systems, ⟨2001, http://www.mpri.lsu.edu/bookindex.html⟩, (accessed July 2, 2010).
31. N.-J. Nilsson, Decision trees, Chapter 6 of introduction to machine learning, ⟨1996, http://ai.stanford.edu/people/nilsson/mlbook.html⟩ (accessed July 2, 2010).
32. P. Parpas, B. Rustem, and E.-N. Pistikopoulos, "Linearly constrained global optimization and stochastic differential equations," *J. Global Optim.* **36**(2), 191–217 (1996).
33. Scharstein and Szeliski, "High-Accuracy Stereo Depth Maps Using Structured Light," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR03),*, Madison, USA, vol. 1, pp. 195-202. IEEE, Washington, DC (June 2003).
34. D. Reid, "An algorithm for tracking multiple targets," *IEEE Trans. Autom. Control* **24**(6), 843–854 (1979).

Tarek Yahiaoui electronic engineer received his PhD degree from the University of Lille, France in 2007 in the field of computer science. He is a researcher in the field of image processing applied to safety and security in public transport. He has several publications mainly about automated people counting and stereo matching. He is currently working in the field of image processing at LIFL Laboratory. University of Lille, France.

Louahdi Khoudour received a degree in applied mathematics from the University of Toulouse in 1992 and a Master Degree in Computer Science from the University of Toulouse in 1993. He then obtained a PhD in Control and Computer Engineering from the University of Lille in 1996. In 2006, he obtained the Leading research degree (Director of research) in Physical Sciences from the University of Paris. He is currently a researcher at INRETS (French National Institute on Transport and Safety Research). Since 1997, he has supervised 6 Phd students (3 completed and 3 ongoing) in the field of computer vision applied to safety and security in public transport. He has been in charge of various European projects, such as Cromatica, Prismatica, Boss, Selcat, Securemetro, PANsafer dealing with safety and security aspects in guided transport systems. His main competencies are video surveillance and image processing applied to safety in public transport. He is author or co-author of around 20 papers in journals, several chapters in books, 50 international conference papers and several grants.

Cyril Meurie received his PhD in Computer Science, from University of Caen Basse-Normandie (Caen) France in 2005. From 2006 to 2008, he was post-doctoral researcher with the Electronic, Waves and Signal Processing Research Laboratory for Transport (LEOST) of the French National Institute for Transport and Safety Research (INRETS). He participated to the European project BOSS (On Board Wireless Secured Video Surveillance) and actually to the French project PANsafer. Since 2008, he is an associate professor with the Systems and Transportation Laboratory (University of Technology of Belfort-Montbeliard). His research interests focus on image segmentation and classification techniques for color and textured images (multi-scale and morphological methods), stereovision approaches, localization and autonomous navigation for intelligent vehicles.