An Embedded System for Counting Passengers in Public Transportation Vehicles

Nicola Bernini, Luca Bombini, Michele Buzzoni, Pietro Cerri, and Paolo Grisleri
Dipartimento di Ingegneria dell'Informazione
University of Parma, ITALY
email: bernini,bombini,buzzoni,cerri,grisleri@ce.unipr.it

***Abstract*—This article describes a system for people counting conceived for public transportation vehicles. The underlying idea is to monitor the number of passengers getting in or out public transportation means like buses and metros over time hence computing reliable estimations in order to improve vehicle's door control. A stereo vision system is presented, it has been developed considering its future installation over bus doors; a feature based people counting algorithm and an object tracking system are used to count people getting in or out of a specific region of interest. The system here described will be installed and tested on an Ivecobus Citelis vehicle in the framework of the Italian Industria 2015 Ecoautobus initiative.**

***Keywords*—Industria2015, Stereo Vision, People Counting, People Detection**

I. Introduction

A. General Introduction

People counting is a task attracting a great interest in several fields: the most important ones concern safety related issues but it also proves useful for economic reasons (specifically regarding marketing). Presently the most common approaches for people counting in the public transportation context rely on mechanical interaction, like in case of turnstiles. This kind of systems show important limitations like slowness and a continuous maintenance need (because of their mechanical nature).

In this article it is presented an unobtrusive computer vision based people counting system along with its application on board of a bus.

The advantages that such a system could bring can be numerous, starting from safety because it could be exploited to avoid overcrowding on the public mean of transport. Another important application could be related to the public transportation service efficiency improvement, as the system would provide real-time data related to the service usage thus allowing to fine tune it according to the needs. Finally, it could be possible to exploit such a system to optimize the doors control in order to reduce the mechanical parts wear.

The approach presented in this work regards the public transport environment as illustrated in Figure 1. It is different with respect to other computer vision based people counting systems based on a stereo couple in zenithal position because it has been developed and tested to optimize it for specific applications in the public transportation context, achieving real time performance while keeping the cost low. The test dataset have been recorded with specific attention to the abovementioned particular scenario: people wearing hats and large backpacks have been used along with specific light conditions for the scene.

B. Different Technologies

Many different technologies are used for people counting.

1) Profile measuring technologies: Sonar is often used as a supporting technology in sensor fusion solutions however its noise sensitivity makes it unsuitable for a stand-alone solution even if its precision is proving useful when properly fused with other technologies.

Laser scanner is the main technology of this category: by using it, an accurate profile of the surrounding can be obtained. The main concern with respect to this kind of approach remains its high cost.

2) Passive Infrared Technologies: IR (Infrared) sensors are cost effective and have proven useful for surveillance solutions however their low resolution makes them inappropriate for people counting applications especially for rather crowded environments, as observed in [1].

IR cameras are a more expensive but also provide a higher resolution technology that could prove beneficial for people detection in real environments.

3) 3D Sensors: Different approaches have been investigated to acquire information about the depth of the image points.

The *structured light* approach consists in the projection of a specific light pattern on an object so that its geometry will modify it, thus with a camera it is possible to measure this deformation inferring the depth information. A famous commercial product exploiting this strategy is the Microsoft Kinect. This device has a rather low cost but presents also serious limitations in terms of its limited usage (because of the used light nature, it does not work outside) and of the high computational cost needed to process its output.

A stereo pair allows the computation of the disparity image, as described in [2] and in [3], furthermore the availability of two points of view makes the system more tolerant to different challenges e.g. occlusions, fast illumination conditions change...With this technology it's easier to distinguish ambiguous problematic cases.

To sum up, this kind of technology introduces the following advantages

(a) (b) (c)

Fig. 1. The Ecobus Project Images: (*a*) The Bus, (*b*) The Computer (prototype), (*c*) Stereo Couple.

- Cost effective way to compute depth information about an image
- Passive technology: no interference with other methods
- Support for different features like: object tracking, recognition, classifications...
- Easy integration

C. People Counting Algorithms

People Counting Algorithms can be grouped in two major categories

- People detection
- Features based

The people detection algorithms goal is to identify every person appearing in the image, according to a certain set of criteria. Once a person has been identified, the people counter gets incremented and the person is tracked along his motion to avoid counting it again in future iterations.

This kind of approaches works well in low crowding conditions, when people in the image are easily recognizable by means of unambiguous elements like the head but works poorly in presence of high crowding conditions when mutual occlusions happen very often.

The feature detection algorithms are best suited for cluttered scenes as they are able to statistically model a crowded environment working on features like edge density, edge orientation, amount of moving pixels, amount of fractal objects, blob size, ...

This kind of approach is in general used more because of its reduced sensitivity to the abovementioned problems of occlusions, typical values in real scenario are shown in [4].

1) People Detection Algorithms: People detection is usually performed by means of model matching, especially concerning unambiguous part of the body like the head. Indeed head detection is one of the most common strategies because important assumptions can be made regarding the head shape and height from the ground level.

This kind of approach works particularly well with certain setups, like placing the stereo couple in zenithal position thus reducing significantly the probability of occlusions, conversely as it happens when the camera is in a general position: in these situations, the occlusions probability is non negligible and a proper method to address them is needed.

A solution well known in literature for head counting has been developed at the **University of Amsterdam** and is presented in [5]. It is based on a stereo pair mounted in zenithal position over a certain passage. Computing the depth information, it is possible to easily discriminate between the background (the floor) and the foreground which contains, in addition to objects, people. To distinguish a person from other foreground objects, the head is searched and the most common approach relies on the disparity point cloud fitting with a spherical model (because of its similarity with the human head).

This kind of strategy alone, suffers false positives due to other spherical objects in the scene (such as spherical fruits, soccer balls...) however other information can be considered to refine the selection like the height of candidate sphere from the ground level, its dimension...

To evaluate the performance, three kind of tests have been carried out, varying parameters like the **distance among people** which is directly related to the *crowding level* and the **homogeneity** of movements.

First a **standard walking test** is performed, consisting of: the application of the people counting system to a group of well distanced people walking along the same direction.

Then a **cross walking test** is performed, consisting of: two groups in which people are always distanced by at least one meter but they are walking in opposite directions.

Finally a **crowded walking test** is performed, consisting of: only one group of people moving coherently in the same direction, similarly to the standard walking test, but the people are one next to the other, leading to a crowded situation.

Each test is evaluated in terms of **precision**, it is to say *the ability to reduce false positives*, and **recall**, that is the *ability to reduce false negatives*.

Another interesting solution has been developed by the **University of Cluj-Napoca** and is presented in [6].

This algorithm proceeds on two paths at the same time.

An edge detection and a non local means filter is applied to a copy of the image coming from the left camera.

At the same time, the unprocessed images coming from both cameras are used to compute a disparity image and the generated point cloud is passed to an object detector, then a foreground extraction is performed, the extracted objects are labeled and a border refinement, using the initial left image, is carried out in order to have a properly surrounded pedestrian candidate. The three candidates are then compared using template matching techniques with pedestrian models.

This approach is quite sensitive to the image quality: starting from a grayscale image, false discontinuities could be generated as a consequence of carrying out differentiation and thresholding as stated in [7] and they could affect the global performance. This kind of challenge can be handled using a high resolution color camera.

2) Feature based people counting: Model matching people detection proves particularly effective in absence of occlusions and when the model is very unambiguous, but these assumptions could not hold well in specific contexts characterized by high crowdedness and heterogeneity. In this kind of situations, feature based strategies could provide better performance.

Edges detection based approaches have been gathering attention for a long time [8] since the information conveyed is high. Some evolutions consisted of combining this source of information with other ones ranging from blob size [9] to the Minkowski fractal dimension [10].

In order to manage the complexity of a crowd some approaches rely on multiple point of views [11].

II. Methods

A. Hardware and Software Framework

The Ecobus Project solution relies on a stereo vision approach: on every door of the bus, a pair of cameras is mounted in zenithal position in a setup similar to the one described in [12] and in [13].

A Bumblebee2 bb2-03s2m with $12\,\mathrm{cm}$ baseline and $2.5\,\mathrm{mm}$ optics providing 640×480 pixel black and white images has been used. It has been mounted forming an angle of about 10 degrees with respect to the zenithal axis.

The algorithm has been executed on a Intel(R) Core(TM) i7-2600 CPU @ 3.40GHz with 4 cores and 8 GB RAM.

The solution has been implemented using the Vislab proprietary software framework GOLD [14].

B. Algorithm Explanation

The algorithm used essentially relies on features extraction for people identification and object tracking to increase (for people getting in) and decrease (for people getting out) the people counter.

1) Blob Generation: Once images have been acquired from both cameras, a first image pre-processing phase is carried out in order to implement dedistorsion and rectification by means of proper LUT application. A downsampling is then carried out in order to define a subset composed of 1/4 of the original points set for each image, to further proceed. This kind of strategy improves the processing speed without affecting the disparity precision for the selected points. Finally, homologous points are identified and a disparity image is computed, according to the algorithm described in [15] and in [16].

Filtering is then performed on this image, applying a ROI, focusing on a $2\,\mathrm{m} \times 2\,\mathrm{m}$ plane surface, and in terms of height from the ground level, focusing in a region that ranges from $1.5\,\mathrm{m}$ to $2.5\,\mathrm{m}$. Then an occupancy grid made of $4\,\mathrm{cm} \times 4\,\mathrm{cm}$ square cell is computed over the remaining points and a clustering is performed relying on the connected components. The clusters whose dimension is bigger than a certain empirical threshold are drawn as blobs.

A feature extraction operation, performed over corners and edges, is carried out on clusters in order to identify them in the frame sequence and to be able to distinguish new ones from the old ones and their directions of movement. To associate two blobs in two different frames, similar features are searched and if a match is found, the tanslation needed to make them coincide, is computed to estimate the blob motion.

2) Blob interframe association: A fundamental task for the overall algorithm is the *blob comparison in two different consecutive timeframes* aimed at carrying out an interframe association process.

Let's consider a certain blob B_0 found inside a frame in t_0. In the frame in $t_1 > t_0$ all of the n blobs $\{B_{1,i}\}_{i=1,\dots,n}$ are compared with B_0 according to a certain feature based similarity measure.

The similarity function is indicated by $s(B', B)$ with B, B' blobs taken from 2 different timeframes. The similarity criteria rely on the percentage of common features shared between the 2 blobs.

At the end of this step a **similarity matrix** M_{n_0,n_1} has been computed, associating the n_0 blobs related to the timeframe in t_0 and the n_1 blobs related to the timeframe t_1.

3) Cluster Fusion: It is rather common for the clustering module to perform suboptimally, because of missing subcluster fusions leading to a missing merge among different body parts of the same person. This kind of behavior can significantly affect the overall performance so a proper *cluster fusion module* has been provided in order to execute the missing fusions.

If for a certain reference blob B_0 taken from timeframe in t_0 a set $\{B_{1,i}\}_{i=1,\dots,m}$ of m blobs taken from $t_1 > t_0$ (consecutive timeframe) is found to be very similar, it means that the similarity measure exceeds a certain s_{th_1} threshold, to B_0 thus

$$s(B_{1,i}, B_0) < s_{\text{th}_1} \qquad i = 1, \dots, m$$

this is a *set of fusion candidates*.

A further refinement on this set is then carried out, identifying a subset $\{B_{1,i}\}_{i=1,\dots,m'}$ of $m' < m$ elements that share the minimal amount of features among them, because they are assumed to be complementary parts of an original blob.

Finally, the blobs in $\{B_{1,i}\}_{i=1,\dots,m'}$ are fused so a new blob $B_1^\star$ is generated and its similarity value $s(B_1^\star, B_0)$ is computed.

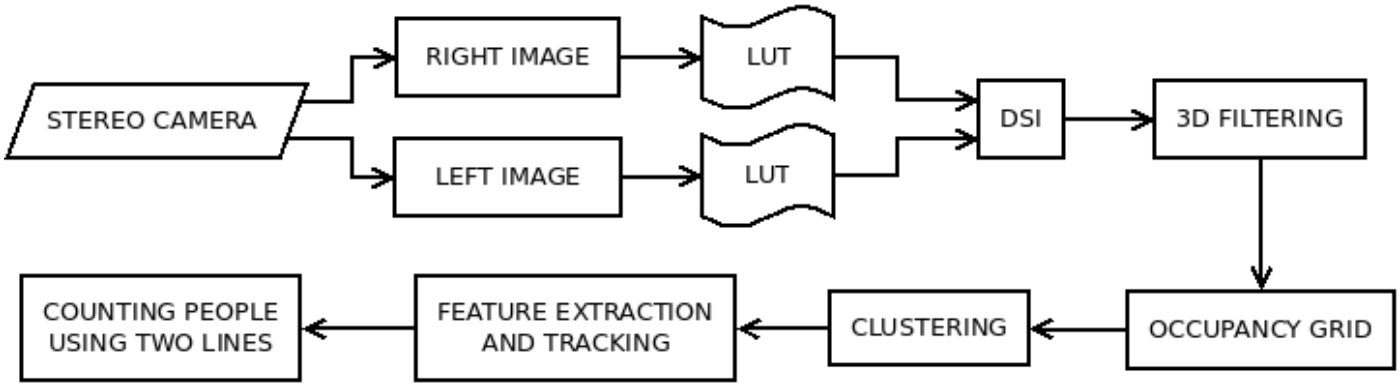
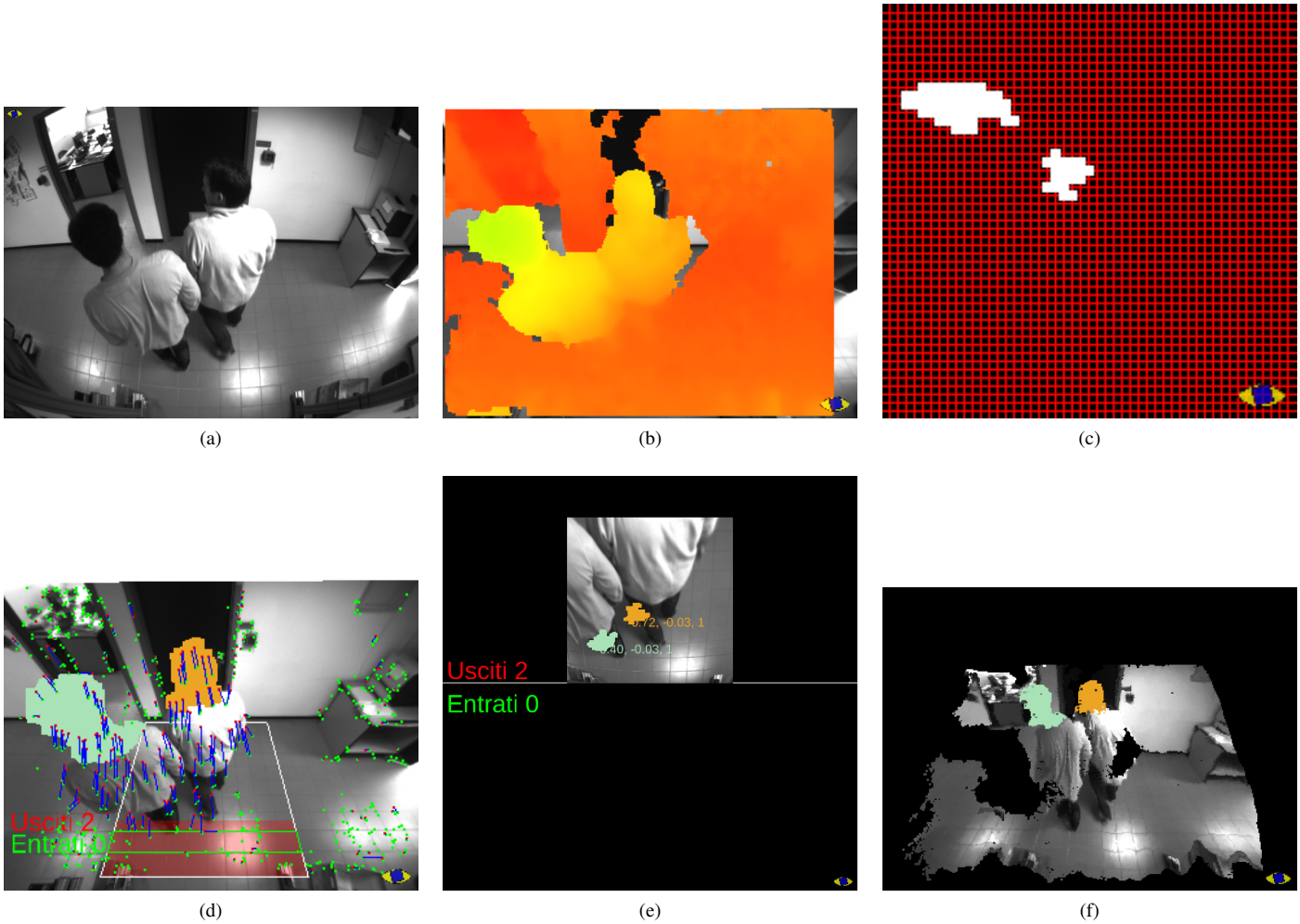


Fig. 2. Algorithm flowchart: the stereo system acquires a pair of images that get dedistorted by means of a proper LUT application, then a Disparity Image is computed and 3D Filtering is performed to select the points inside the pre-defined ROI. Hereafter the Occupancy Grid is computed and a Clustering Process is performed on the result. Features are then extracted out of clusters and a correspondence is established among blobs in different frames, so that to be able to recognize situations where objects get in or out the ROI and to perform object tracking, hence leading to people counting.



(a) (b) (c)

(d) (e) (f)

Fig. 3. Different processing steps: (*a*)input image: (*b*) DSI, (*c*) occupancy grid, (*d*) feature tracking, (*e*) people counting on IPM view, and (*f*) 3D reconstruction. Notr that the occupancy grid adopts a different convention regarding the reference system.

Thus the above mentioned **similarity matrix** is transformed in terms of deletion of the columns related to the fused blobs and creation of the columns related to the result of these fusions, so the matrix becomes $M_{n_0,n_1'}$ with $n_1' < n_1$.

4) Entering and Exiting: If no elements have entered or exited the image ROI, assuming that the cluster fusion module has worked properly, the similarity matrix is expected to be square. To associate a blob in t_0 to a blob in $t_1 > t_0$ the

following maximization problem is solved

$$B_1 = B_{1,i^\star} \qquad i^\star = \arg\max_i \{s_i = s(B_{1,i}, B_0) \text{ if } s_i > s_{\text{th}_2}\}$$

Tuning this threshold is important to avoid incorrect associations when the number of elements exiting is equal to the number of elements entering.

In presence of a net amount of objects entering the sensitive area, the matrix becomes rectangular with $n_1 > n_0'$ and some columns show low similarity values, under the s_{th_2} if properly tuned, in each of their elements.

In presence of a net amount of objects exiting the sensitive area, the matrix becomes rectangular with $n_0 > n_1'$ and some rows show lower similarity values, under the s_{th_2} if properly tuned, in each of their elements.

In order to update the people counter properly, the following criteria are applied.

First a *relevant movement direction* is identified: people moving along that direction in one way are entering and people moving along the same direction in the opposite way are exiting.

As parameters for the algorithm, two threshold positions along the above mentioned movement direction are then initially defined inside the Image ROI: let's call them P_1, P_2 where conventionally it's $P_1 < P_2$

Let's define as $P(B,t)$ the position along the relevant movement direction for a generic blob B at the generic time t

The condition related to the "element getting inside" event, is the following one: if for some B it is observed that

$$P(B,t_0) < P_1 \text{ and } \exists t_1 > t_0, P(B,t_1) > P_2$$

the people counter is increased.

The condition related to the "element getting outside" event, is the following one: if for some B it is observed that

$$P(B,t_0) > P_2 \text{ and } \exists t_1 > t_0, P(B,t_1) < P_1$$

the people counter is decreased.

III. Discussion

The setup used for the tests, consists of the single bus door, mounted at ground level, not on board of a vehicle. The algorithm has been tested using three different video sequences: one inside, 10 hours long (about 360000 frames); two outside, one in bright light conditions, 6 hours long (216000 frames) and one in dim light conditions, 4 hours long (144000 frames). No specific assumptions have been made regarding the human figure: in the sequence it is possible to find people with hats and backpacks, group of people, wheelchairs and strollers.

	Indoor	Outdoor (Bright)	Outdoor (Dim)
IN GT/Counted	542/531	211/204	183/160
OUT GT/Counted	536/528	226/215	165/148
False Positives IN/OUT	5/4	2/1	4/4
False Negatives IN/OUT	16/12	9/12	27/21

TABLE I. System performance.

The algorithm proves very precise in terms of person detection as suggested by the data in table I.

After a more detailed analysis has been carried out, it has been observed that the main cause for the above mentioned errors is the lack of features that affects negatively the inter-frame blob association thus leading to a global performance decrease. This kind of phenomenon is particularly pronounced in dim light conditions. It has been observed that the system exceeds the cluster fusion activity and tends to group different people in a single blob, especially in high crowding conditions.

IV. Conclusion

A. System Performance

The algorithm has been set up to work at 10 frames per second, in line with the image acquisition time: it always succeeds in finishing its processing in the $100\,\text{ms}$ time slice available.

The tests have shown good algorithm performances in the presented implementation, along with important suggestions to still enhance them.

In figure 4 some results are shown, (a), (b), (d), and (e) show succesful results in complex situations like (a) two embraced people, (b) a man with a crate, (d) the person is detected regardless of the hat he is wearing, or (e) the detection of two persons.

In (c), a false negative is shown: only one of the two heads in the scene is detected because of the low number of features that have been computed.

As a further development, the minimum height from the ground level for the disparity image ROI, could be modified to values lower than $1.5\,\text{m}$ thus considering children too.

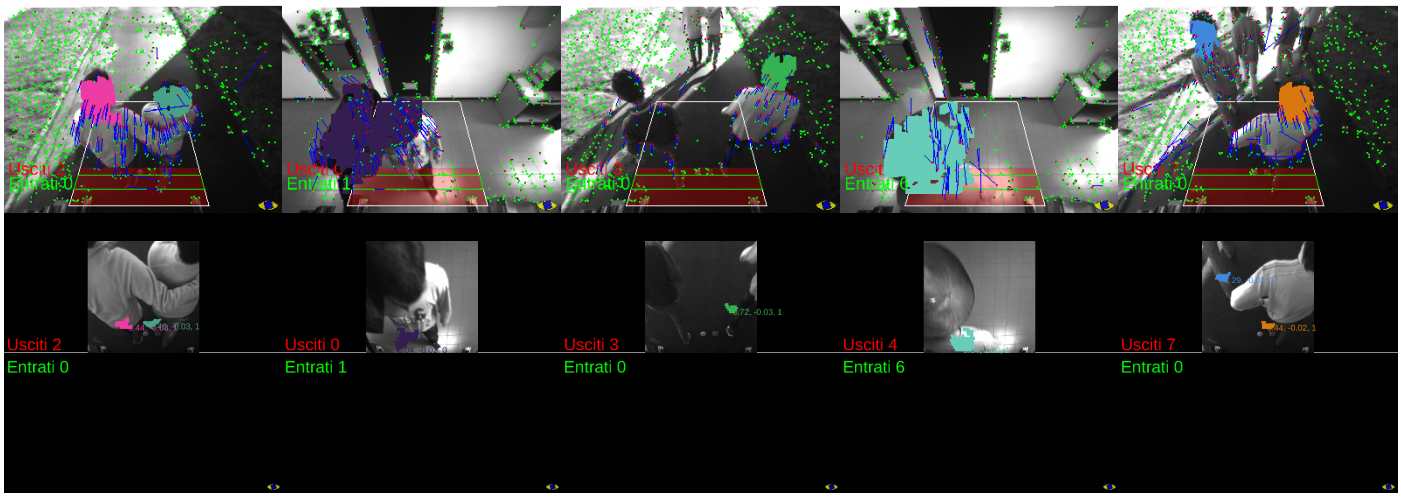
It is necessary to consider that with this kind of extension a wide range of non human elements will be evaluated by the system and in addition to the extra computational cost required, a negative impact in terms of precision and recall could occur because of the problematic discrimination, in the context of the present solution, among children and some objects like backpacks.

B. Further Developments

A first improvement could consist in considering different features extraction strategies. With respect to the ROI lower level limit extension, the development of a template matching module can be considered so to improve the discrimination capabilities regarding human and non human elements, but its performance could be affected by the zenithal point of view. A frontal camera, or camera pair, associated to the original stereo couple, can surely be a very valuable source of information.

Acknowledgment

The work described in this paper has been carried out in the framework of the Industria2015 Ecoautobus project funded by the Italian Government.



(a) (b) (c) (d) (e)

Fig. 4. Some results, in the upper part the original images, in the lower part the results (the counter shown is not related to the single frame but to the history hence it might be misleading): (*a*) embraced people, (*b*) crate on the shoulder, (*c*) left person head not detected, (*d*) people with a hat, (*e*) two heads detected.

REFERENCES

[1] D. Beymer and K. Konolige, "Real-time Tracking of Multiple People using Continuous Detection," in *Procs. Intl. Conf. on Computer Vision*, Kerkyra, 1999.

[2] H. Sunyoto, W. Van der Mark, and D. M. Gavrila, "A comparative study of fast dense stereo vision algorithms," in *Intelligent Vehicles Symposium, 2004 IEEE*. IEEE, 2004, pp. 319–324.

[3] W. Van Der Mark and D. M. Gavrila, "Real-time dense stereo for intelligent vehicles," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 7, no. 1, pp. 38–50, 2006.

[4] G. Mori, S. Belongie, and J. Malik, "Efficient shape matching using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1832–1837, 2005.

[5] T. van Oosterhout, S. Bakkes, and B. J. Kröse, "Head detection in stereo data for people counting and segmentation." in *VISAPP*, 2011, pp. 620–625.

[6] I. Giosan, S. Nedevschi, and S. Bota, "Real time stereo vision based pedestrian detection using full body contours," in *Intelligent Computer Communication and Processing, 2009. ICCP 2009. IEEE 5th International Conference on*. IEEE, 2009, pp. 79–86.

[7] S. T. Birchfield, B. Natarajan, and C. Tomasi, "Correspondence as energy-based segmentation," *Image and Vision Computing*, vol. 25, no. 8, pp. 1329–1340, 2007.

[8] S.-Y. Cho, T. W. Chow, and C.-T. Leung, "A neural-based crowd estimation by hybrid global learning algorithm," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 29, no. 4, pp. 535–541, 1999.

[9] D. Kong, D. Gray, and H. Tao, "A viewpoint invariant approach for crowd counting," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 3. IEEE, 2006, pp. 1187–1190.

[10] A. B. Chan, Z.-S. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–7.

[11] D. B. Yang, H. H. González-Baños, and L. J. Guibas, "Counting people in crowds with a real-time network of simple image sensors," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 122–129.

[12] Z. Qiuyu, T. Li, J. Yiping, and D. Wei-jun, "A novel approach of counting people based on stereovision and dsp," in *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*, vol. 1. IEEE, 2010, pp. 81–84.

[13] K. Terada, D. Yoshida, S. Oe, and J. Yamaguchi, "A counting method of the number of passing people using a stereo camera," in *Industrial Electronics Society, 1999. IECON'99 Proceedings. The 25th Annual Conference of the IEEE*, vol. 3. IEEE, 1999, pp. 1318–1323.

[14] M. Bertozzi, L. Bombini, A. Broggi, P. Cerri, P. Grisleri, and P. Zani, "GOLD: A framework for developing intelligent-vehicle vision applications," *IEEE Intelligent Systems*, vol. 23, no. 1, pp. 69–71, Jan.–Feb. 2008.

[15] A. Broggi, M. Buzzoni, M. Felisa, and P. Zani, "Stereo obstacle detection in challenging environments: the VIAC experience," in *Procs. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, San Francisco, California, USA, Sep. 2011, pp. 1599–1604.

[16] N. Morales, G. Camellini, M. Felisa, P. Grisleri, and P. Zani, "Performance Analysis of Stereo Reconstruction Algorithms," in *Procs. IEEE Intl. Conf. on Intelligent Transportation Systems*, The Hague, The Netherlands, Oct. 2013, pp. 1298–1303.