Low-level Pedestrian Detection by means of Visible and Far Infra-red Tetra-vision

M. Bertozzi, A. Broggi, M. Felisa, G. Vezzoni
Dipartimento di Ingegneria dell'Informazione
Università di Parma
Parma, I-43100, Italy
{bertozzi,broggi,felisa,vezzoni}@ce.unipr.it

M. Del Rose
Vetronics Research Center
U.S.Army TARDEC
Warren, MI, U.S.A.
DelRoseM@tacom.army.mil

***Abstract*— This article presents a tetra-vision (4 cameras) system for the detection of pedestrians by the means of the simultaneous use of one far infra-red and one visible cameras stereo pairs. The main idea is to exploit both the advantages of far infra-red and visible cameras trying at the same time to benefit from the use of each system. Initially, the two stereo flows are independently processed, then the results are fused together. The final result of this low-level processing is a list of obstacles that have a shape and a size compatible with the presence of a potential pedestrian. In addition, the system is able to remove the background from the detected obstacles to simplify a possible further high level processing.**

The developed system has been installed on an experimental vehicle and preliminarily tested in different situations.

I. Introduction

The detection of obstacles and pedestrians is one of the most active research targets for public, commercial, and governments organizations.

I particular, the U. S. Army is actively developing obstacle detection for mule operations, path following and intent based anti-tamper surveillance systems for its robotic vehicles safety [8,9,15].

Nevertheless, the potential field of use of such technology is larger than U. S. Army necessities: surveillance systems, driver assistance systems, and intelligent systems for autonomous or semi-autonomous driving represents few of the many areas that need to detect the presence of persons in order to take appropriate actions like avoiding it or enabling safety countermeasures.

Unfortunately, the detection of pedestrians is a really challenging task and problem increases in difficulty when considering the movement of the sensors, uncontrolled outdoor environments, and variations in pedestrian's appearance and pose.

In the last years, many different approaches to solve this problem have been tested. Some use LADAR or laser scanners to retrieve a 3D map of the terrain and detect pedestrians [10–12], another uses ultrasonic sensors to determine the reflection of pedestrians [1]. Radar is also popular for detecting pedestrians similar to ultrasonic sensors; by measuring the reflection of possible targets and determining if they are pedestrians or not [16,18].

A natural choice for a pedestrian detection sensor is vision because it is based on how people perceive humans (through visual cues). In particular, for the U. S. Army vision based detection is important since cameras are non-evasive sensors. Within this problem set there are monocular vision systems [21,23,26] or stereo vision systems [3,6,13,20,22,24].

Recently also infra-red technologies (both far and near infra-red based) have made their appearance in the pedestrian detection arena [2,4,7,19,25], thanks to the decreasing cost of infra-red technology.

In many situations, infra-red technology presents many advantages with respect to visible-based systems. In particular, in the far infra-red domain warm obstacles or pedestrians are warmer than the environment and, therefore, brighter than the background. Anyway, far infra-red cameras fail the detection of pedestrians in hot or sunny weather namely when pedestrians are not warmer than the background.

The following presents a pedestrian detection system based on the contemporary use of a visible and a far infra-red stereo systems in order to exploit the benefits of both approaches.

This paper is organized as follows: section II introduces all parts of the algorithm and section III presents the results of this approach and the performance of the system. Section IV summarizes and concludes the paper.

II. The Approach

Far infra-red systems and visible cameras have been widely used for the detection of obstacles. The choice between infra-red and daylight technologies generally depends on the specific use cases and on costs vs benefits analysis.

Far infra-red sensors, in fact, sense the thermal features of the scene. In the far infra-red domain the image of an object depends on its temperature and the amount of heat it emits and is not affected by illumination changes.

Therefore, far infra-red cameras are suitable for the detection of objects warmer (or colder) than the background (pedestrians, moving vehicles,...), since they are sufficiently contrasted with respect to the background.

In addition, images acquired by far infra-red cameras do not depend on illumination; they can be used in day or night-time with little difference and are not affected by the presence of shadows. Also the lack of textures and colors can be exploited to filter out small noisy details.

Unfortunately, strong sun heating or high temperature conditions can increase the background thermal impact in the far

infra-red images and can also introduce additional textures due to the different thermal behavior of different materials.

In the specific case of pedestrian detection, clothes greatly affect the thermal footprint of a person making the human shape detection challenging.

Conversely, the detection of pedestrians in the visible domain is often more difficult due to the presence of small details, shadows, and changes in the luminance or sharpness of acquired images due to different illumination conditions.

Anyway, the presence of details can be considered as noise in an initial phase of a detection system where shape is more often used as a preliminary detector, but is of paramount importance for distinguishing between pedestrians and human-shaped objects.

In the following a system based on the simultaneous use of two infra-red and visible cameras stereo pairs is described. The main idea is to exploit both the advantages of far infra-red and visible cameras trying at the same time to cope with the deficiencies of each system.

A. The algorithm

Dealing with two stereo systems that work in different domains, is a bit tricky since the two stereo systems feature different points of view, different angular apertures, and –generally– different frame rates and resolutions. In addition, the system can not rely on a search for homologous points, since it is very difficult to match features across different spectral domains.

Many image registration strategies have been proposed [14], but unfortunately they require annotation of homologous points or an a-priori knowledge of the environment.

Conversely, the proposed system independently processes the two stereo flows and then fuses the results that come from the different domains. The main steps of the algorithm are:

1) Road slope detection: many vision-based systems for intelligent vehicles rely on the assumption of moving on a flat road or –more realistically– on a road with a smoothly varying slope. Anyway, this assumption can lead to errors in computing pedestrians parameters or even to misdetections. Especially for vision systems installed on moving vehicles, pitch and roll can in fact void this assumption.

A v-disparity approach is used for computing the actual road slope, exploiting the presence of stereo vision systems [5,17].

A Sobel filtering is used to enhance images features, namely edges. Then, a correlation is computed for different offset values (*disparities*), for each pair (left and right) of rows of the Sobel images (thanks to the specific setup of the stereo pair, the same rows in the stereo images are epipolar lines). This process is performed both for visible and infra-red images; the result is a new image (the *correlation map*) encoding correlation values (see figure 1). The brighter the pixel the higher the match quality.

It can be noticed that for the visible images, the ground components produce a slanted line that can give information about the road cross section. Conversely, due to the bare presence of ground edges in infra-red images, the correlation map does not permit to recover road data.

Fig. 1. V-disparity computation in visible and infra-red domains: (*left*) original images, (*center*) Sobel images, and (*right*) correlation map.

Therefore, only the visible domain is used to compute the ground slope; anyway, this information can be used in the following process also for the infra-red domain thanks to the knowledge of relative calibration of all four cameras.

2) Independent obstacle detection: during this step of image processing, the two stereo flows are independently processed to detect obstacles. A *disparity space image* process is used to detect the presence of obstacles.

The right images of each flow are subdivided into 3×8 pixels regions and their corresponding regions are searched for into the left images. The 3×8 size for search regions is due to the fact that a pedestrian shape is generally characterized by strong vertical features. The search for a homologous region is limited to the same row of the left image, since the optical axis of the two stereo systems are parallel and thus two corresponding rows in the two images are epipolar lines. Moreover, calibration information permit to further bound the search area, reducing both the required computational burden and the risk of wrong matches. For each region in the reference image, the best matching region in the other image is considered and the disparity between their coordinates is computed.

The match is performed using the following correlation formula:

$$C = \frac{\sum_{i=1}^{3}\sum_{j=1}^{8}(L_{i,j} \times R_{i,j})}{max\left(\sum_{i=1}^{3}\sum_{j=1}^{8}(L_{i,j})^2, \sum_{i=1}^{3}\sum_{j=1}^{8}(R_{i,j})^2\right)} \qquad (1)$$

where (i, j) are the coordinates of each pixel into the 3×8 regions. The result is a new image, the disparity space image (DSI), in which each pixel encodes the disparity value. Figure 2.a shows an example of DSIs computed both for the visible and infra-red domains; the colors encode the disparity value: bright colors correspond to small disparities, while dark colors correspond to high disparity values. It can be noticed that only obstacles feature large clusters of pixels that have similar disparity values, conversely background features

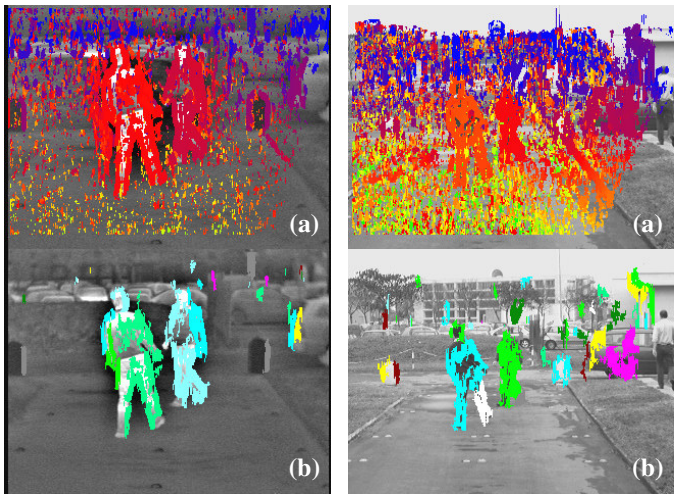

Fig. 2. DSI: (*a*) the disparity images for visible and infra-red domains (the darker the color, the smaller the disparity) and (*b*) detected obstacles regions labelled using different colors.

present colors that evolve from bright to dark when moving from the bottom part of the image towards the top.

Therefore, an aggregation step on the DSI is used to detect the presence of obstacles: large areas featuring a similar disparity are marked as obstacles. Initially, DSI components due to the background are removed: each column of the DSI is considered, and portions of each column that present variable disparity values are removed. Not all surviving cluster of pixels are considered as obstacles. In fact, size, distance, and height constraints are used to further reduce their number. Figure 2.b shows resulting clusters with different colors.

3) Merge and refinement step: each region produced by the segmentation process described in previous section is marked using a bounding box (see figure 3.a).

Unfortunately, different bounding boxes often belong to the same obstacle. Therefore a merging process is mandatory for obstacle detection. Two bounding boxes are fused to compose a larger bounding box when they are sufficiently close in the real world coordinates. Two different proximity parameters are used:

Disparity: only bounding boxes that feature a similar disparity can be merged, since the disparity encodes the distance of bounding boxes from the vision systems and thus bounding boxes that belong to the same obstacle should feature a similar disparity.

Transversal distance: in order to be merged, two bounding boxes have to be closer than a given threshold distance in the real world coordinates.

Figure 3 shows an example of the merging process in the infra-red domain; thanks to the simultaneous use of disparity and real world coordinates for the merging, bounding boxes belonging to obstacles that partially overlap are not merged (see figure 4).

After the merging step another filtering phase takes place. Bounding boxes too small, too big, or featuring an aspect ratio

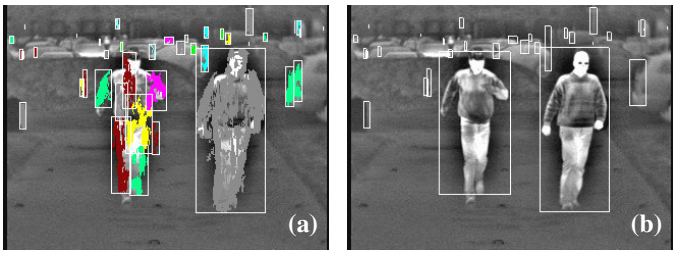

Fig. 3. Bounding boxes generation and merging: (*a*) detected regions enclosed using bounding boxes and (*b*) resulting bounding boxes after the fusion phase.

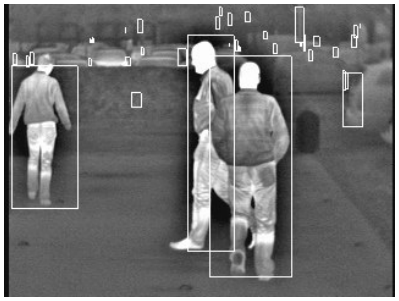

Fig. 4. Overlapping obstacles are correctly detected and enclosed in different bounding boxes.

incompatible with a human shape are discarded.

Finally, the size of resulting bounding boxes is refined. In fact thanks to the knowledge of the road slope computed as described in section II-A.1, it is possible to compute the point of contact between each object framed by a bounding box and the ground. Thus, the bounding boxes' bottoms are stretched to the ground [2] (see figure 5). This allows to include legs when they are misdetected and, at the same time, to remove portions of the background below the human shape when it has been incorrectly included in a bounding box.

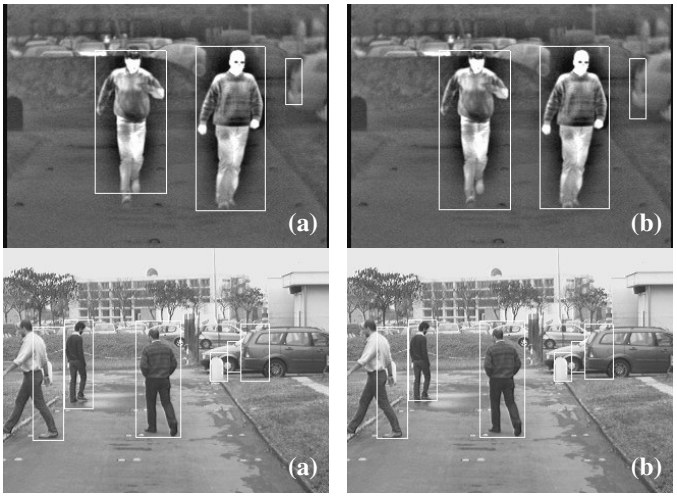

Fig. 5. Two different cases of bounding boxes refinement for the two stereo flows: (*a*) bounding boxes before and (*b*) after the bottom refinement.

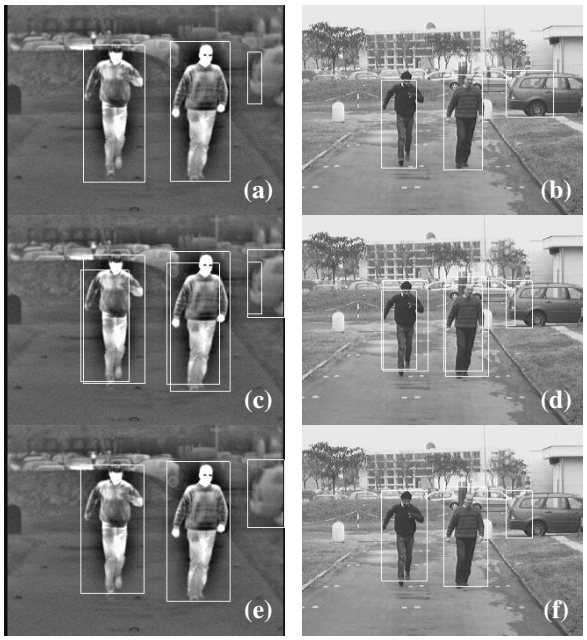

Fig. 6. Bounding boxes registration and fusion. Bounding boxes computed in the (*a*) visible and (*b*) infra-red domains and their union in the (*c*) infra-red and (*d*) visible domains. (*e*) and (*f*) bounding boxes fusion.

4) Bounding Boxes registration: at the end of the previous stage two lists of bounding boxes have been produced: one for the obstacles detected in the visible domain and another one for the obstacles detected in the infra-red images.

Due to the different extrinsic and intrinsic calibration parameters of the two stereo systems, a registration phase in required. In fact, even bounding boxes that correspond to the same obstacle in the two lists are differently sized and positioned. Therefore, in order to fuse the results obtained working on the two spectral domains, it is necessary to compute the position and size of bounding boxes of each list in the other domain. Moreover, the different field of view of the two stereo systems has to be taken in account; in this case, visible cameras have a larger field of view and not all detected objects have a correspondent in the infra-red images.

Thanks to the knowledge of the calibration data, the bounding boxes computed in the visible domain are resized and repositioned according to the infra-red vision system setup and added to the bounding boxes detected in the infra-red domain. Bounding boxes that are completely or partially out of the field of view of the infra-red stereo system are removed or cropped.

The same process is performed to compute size and position of bounding boxes detected in the infra-red images towards the visible domain.

5) Cross-domain fusion of results: since most of the obstacles in front of the vision system are correctly detected both in the visible and in the infra-red domains, the two separate processings can produce two different bounding boxes that correspond to the same obstacle. Moreover, during the previous step, the two lists of bounding boxes have been merged; therefore in each list two bounding boxes can refer to the same object and a merge step is needed. The merging is operated using a bounding box that encloses the two bounding boxes to be fused together. To determine if two bounding boxes refer to the same object, the percentage of overlapping in the images and their distance in world coordinates are considered.

The merging is only performed on the list of bounding boxes in the infra-red domain and is replicated on the homologous boxes in the visible domain (see figure 6.e and 6.f).

At the end of this stage the two lists of bounding boxes contain the same number of bounding boxes and each bounding box in a list corresponds to its homologous one in the other list.

6) Background removal: the DSI images can be further exploited to remove the background from the area enclosed in each bounding box. The DSI components used to compute each bounding box, in fact, have been built thanks to the features that belong to the obstacle and therefore can be used as a mask for removing other nuisances.

Both features detected in the infra-red and visible images are used; they are resized to fit the bounding box size and are fused in a single mask using a OR operator. A morphological expansion is used to fill small holes in this mask.

The final result is obtained ANDing the mask and the original images (see figure 7).

III. Results

Figure 8 shows the results of the whole process in different situations; with or without daylight, rural and urban scenarios, presence of few or many obstacles/pedestrians... It can be noticed that the system is able to detect obstacles even when they do not appear in the images acquired using the daylight cameras or when not sufficiently contrasted with respect to the background.

Performance of the system has been measured using a pre-recorded sequence of about 5000 images with ground truth information about the presence of pedestrians. Since the system is a low-level preprocessing, whose final aim is the detection of possible pedestrians, only the percentage of correct detection is meaningful. The system has proven to be able to detect more than 95% of pedestrians up to 45 m and more than 89% up to 75 m. In order to discriminate pedestrians amongst all the detected obstacles, a higher level processing of each bounding box is needed. On this purpose two approaches based on active contours and matching with pedestrian models are under development. Moreover, as the system knows the dimensions and distances of every obstacle, it is already able to filter bounding boxes incompatible with pedestrian size.

IV. Conclusions

In this paper a tetra-vision system aimed at the detection of pedestrians using two infrared and daylight stereo systems has been discussed.

The two stereo flows are independently processed and the two results are fused together to retain the advantages of both visible and infra-red systems.

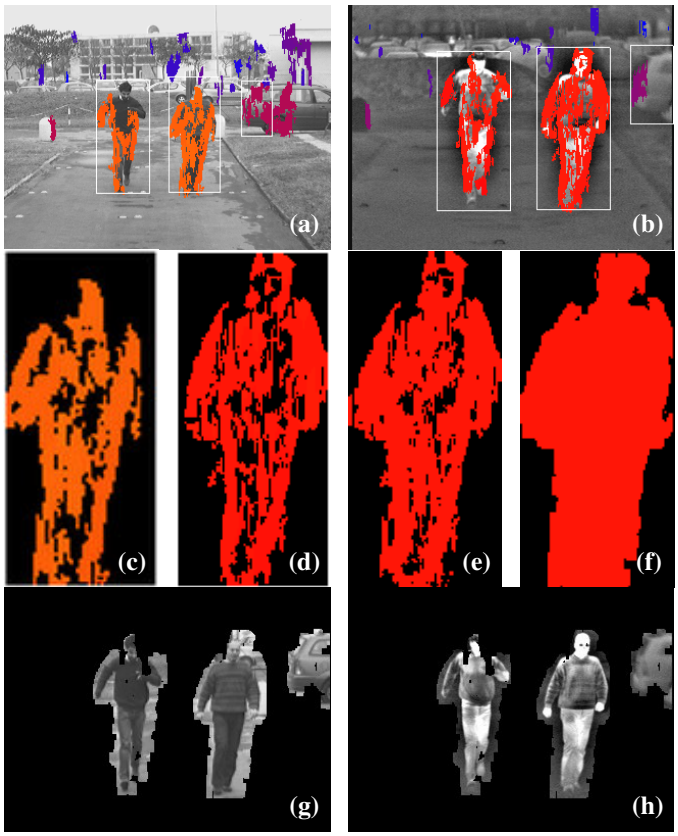

Fig. 7. Removal of background for detected obstacles: (*a*) and (*b*) original images, (*c*) and (*d*) disparity components of right pedestrian in visible and infra-red domains respectively, (*e*) disparity components after their fusion, (*f*) mask for background removal, and (*g*) and (*h*) final result after background removal. Image (*g*) has been cropped to match the field of view of (*h*).

Experimental results demonstrated that this approach is very promising since the system is able to detect pedestrians when they are not viewable by the daylight cameras or when they are not warmer than the background and therefore are not sufficiently contrasted in the infra-red domain. Correct detection percentage is high and the system has proven to work reliably even when pedestrians are partly occluded. Currently, a higher level system aimed at distinguishing between pedestrians and obstacles is under development.

Neither temporal correlation, nor motion cues are currently used for the processing.

ACKNOWLEDGMENT

This work has been supported by the European Research Office of the U. S. Army under contract number N62558-05-P-0380.

REFERENCES

[1] NIT Phase II: Evaluation of Non-Intrusive Technologies for Traffic Detection. Tech. Report No. 3683, Mn DOT Research Report, 2002.

[2] M. Bertozzi, A. Broggi, M. Del Rose, and A. Lasagni. Infrared Stereo Vision-based Human Shape Detection. In *Procs. IEEE Intelligent Vehicles Symposium 2005*, pages 23–28, Las Vegas, USA, June 2005.

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

[4] B. Bhanu and J. Han. Kinematics-based Human Motion Analysis in Infrared Sequences. In *Procs. IEEE Intl. Workshop on Applications of Computer Vision*, Orlando, USA, 2002.

[5] A. Broggi, C. Caraffi, R. I. Fedriga, and P. Grisleri. Obstacle Detection with Stereo Vision for off-road Vehicle Navigation. In *Procs. Intl. IEEE Wks. on Machine Vision for Intelligent Vehicles*, San Diego, June 2005.

[6] R. Cutler and L. S. Davis. Robust Real-time Periodic Motion Detection, Analysis and Applications. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):781–796, Aug. 2000.

[7] J. W. Davis and V. Sharma. Robust Background-Subtraction for Person Detection in Thermal Imagery. In *Procs. Intl. IEEE Wks. on Object Tracking and Classification Beyond the Visible Spectrum*, Washington D. C., USA, 2004.

[8] M. Del Rose and P. Frederick. Pedestrian Detection. In *Procs. Intelligent Vehicle Systems Symposium*, Traverse City, USA, 2005.

[9] M. Del Rose, P. Frederick, and J. Reed. Pedestrian Detection for Anti-Tamper Vehicle Protection. In *Procs. Ground Vehicle Survivability Symposium*, Monterey, USA, 2005.

[10] A. Fod, A. Howard, and M. J. Mataric. Laser-Based People Tracking. In *Procs. IEEE Intl. Conf. on Robotics and Automation*, Washington D. C., USA, 2002.

[11] K. C. Frerstenberg, J. Dietmayer, and V. Willhoeft. Pedestrian Recognition in Urban Traffic using Vehicle Based Multilayer Laserscanner. In *Procs. Automobile Engineers Cooperation International Conf.*, Paris, France, 2001.

[12] K. C. Frerstenberg and U. Lages. Pedestrian Detection and Classification by Laserscanners. In *Procs. IEEE Intelligent Vehicles Symposium 2002*, Paris, France, June 2002.

[13] K. Grauman and T. Darrell. Fast Contour Matching Using Approximate Earth Mover's Distance. Technical Report AI Memo, AIM-2003-026, MIT, 2003.

[14] G. Hines, Z. ur Rahman, D. Jobson, and G. Woodell. Multi-image registration for an enhanced vision system, Apr. 2002.

[15] R. Kania, M. Del Rose, and P. Frederick. Autonomous Robotic Following Using Vision Based Techniques. In *Procs. Ground Vehicle Survivability Symposium*, Monterey, USA, 2005.

[16] M. Koltz and H. Rohling. 24 GHz Radar Sensors for Automotive Applications. In *Procs. Intl. Conf. on Microwaves and Radar*, Warsaw, Poland, 2000.

[17] R. Labayrade and D. Aubert. A Single Framework for Vehicle Roll, Pitch, Yaw Estimation and Obstacles Detection by Stereovision. In *Procs. IEEE Intelligent Vehicles Symposium 2003*, pages 31–36, Columbus, USA, June 2003.

[18] S. Milch and M. Behrens. Pedestrian Detection with Radar and Computer Vision. In *Procs. Conf. on Progress in Automobile Lighting*, Darmstadt, Germany, 2001.

[19] H. Nanda and L. Davis. Probabilistic Template Based Pedestrian Detection in Infrared Videos. In *Procs. IEEE Intelligent Vehicles Symposium 2002*, Paris, France, June 2002.

[20] C. Papageorgiou, T. Evgeniou, and T. Poggio. A Trainable Pedestrian Detection System. volume 38, pages 15–33, June 2000.

[21] A. Shashua, Y. Gdalyahu, and G. Hayun. Pedestrian Detection for Driving Assistance Systems: Single-frame Classification and System level Performance. In *Procs. IEEE Intelligent Vehicles Symposium 2004*, Parma, Italy, June 2004.

[22] H. Shimizu and T. Poggie. Direction Estimation of Pedestrian from Multiple Still Images. In *Procs. IEEE Intelligent Vehicles Symposium 2004*, Parma, Italy, June 2004.

[23] G. P. Stein, O. Mano, and A. Shashua. Vision based ACC with a Single Camera: Bounds on Range and Range Rate Accuracy. In *Procs. IEEE Intelligent Vehicles Symposium 2003*, Columbus, USA, June 2003.

[24] S. Tate and Y. Takefuji. Video-based Human Shape Detection Deformable Templates and Neural Network. In *Procs. of Knowledge Engineering System Conf.*, Crema, Italy, 2002.

[25] F. Xu and K. Fujimura. Pedestrian Detection and Tracking with Night Vision. In *Procs. IEEE Intelligent Vehicles Symposium 2002*, Paris, France, June 2002.

[26] L. Zhao. *Dressed Human Modeling, Detection, and Parts Localization*. Ph.D. dissertation, Carnegie Mellon University, 2001.

Fig. 8. Few results in different situations: with few or many obstacles, simple or complex situations, day or night time. Images (*a*) and (*b*) show the resulting list of bounding boxes superimposed onto the original far infra-red and visible images respectively, and (*c*) and (*d*) present the detected obstacles without the background. It can be noticed that the system is able to detect obstacles even when one of the two stereo systems fails the detection.