Stereo Vision-based approaches for Pedestrian Detection

M. Bertozzi E. Binelli A. Broggi

Dip. Ingegneria dell'Informazione
Università di Parma
43100 Parma, ITALY

M. Del Rose

Vetronics Research Center
U.S.Army TARDEC
Warren, MI, U.S.A.

Abstract

This paper describes a system for pedestrian detection in stereo infrared images. The system is based on three different underlying approaches: warm area detection, edge-based detection, and v-disparity computation. Stereo is also used for computing the distance and size of detected objects. A final validation process is performed using head morphological and thermal characteristics. Neither temporal correlation, nor motion cues are used in this processing.

The developed system has been implemented on an experimental vehicle equipped with two infrared camera and preliminarily tested in different situations.

1 Introduction

The second largest source of automotive related injuries or fatalities is due to accidents that involve pedestrians. To mitigate this problem, several researches groups are working on systems for the detection of pedestrians to be installed on moving vehicles. A promising approach is the use of video sensors that do not emit signals and provide a large amount of informations about the scenario. Unfortunately, vision-based pedestrian detection is a challenging task: pedestrians usually wear different and differently colored clothes and often are barely distinguishable from the background. Moreover, pedestrians can wear or carry items like hats, bags, umbrellas, and many others, which give a broad variability to their shape. Additional problems that must be considered are: noise produced by the presence of buildings and human artifacts, moving or parked cars, cycles, road signs, signals, different illumination conditions, obstacles and so on.

Thus, different approaches have been tested. Some use learning machines like neural networks [14] or support vector machines [8], some are based on the detection of specific patterns, texture or motion clues [11, 12], others are stereo vision based [5].

Only recently, thanks to the decreasing cost of infrared (*IR*) cameras, different systems based on the processing of far-infrared images have been presented [7, 4, 13, 6, 3].

This work presents a system based on stereo infrared images for the detection of pedestrians. The system exploits three different approaches for detecting objects: *i*. warm areas detection, *ii*. vertical edges detection, and *iii*. v-disparity approach. The algorithm groups detected objects with similar coordinates, creating a list of areas of attention. Only areas with specific size and aspect-ratio are considered and filtered using head models that encode morphological and thermal characteristics. Although the proposed method works on single frames and performs no tracking, preliminary results have proven to be promising.

This paper is organized as follows: section 2 introduces all parts of the algorithm and section 3 presents the results of this approach and the performances of the system. Section 4 summarizes and concludes the paper.

2 Algorithm description

In the infrared domain the image of an object relates to its temperature and the amount of heat it emits. Generally, the temperature of people is higher than the environment temperature and their heat radiation is sufficiently high compared to the background. Such human shapes appear brighter than the background in infrared images easing a detection process. In fact, a previously developed approach [2] exploits this feature to detect pedestrians.

Unfortunately, pedestrians are not always brighter than the background. during the summer or in hot environment pedestrians are often darker than the surrounding environment. In addition, clothing may mask heat radiation therefore leading systems based on this feature to partly or complete misdetections of pedestrians.

In order to cope with this problem, three different underlying approaches have been developed for pedestrian detection: warm areas detection, vertical edges detection, and v-

0-7695-2372-2/05/$20.00 (c) 2005 IEEE

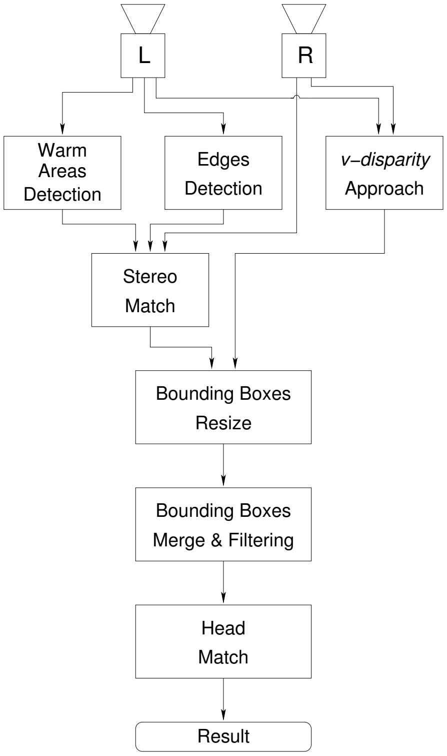

Figure 1: Overall algorithm flow.

disparity approach. The first approach is devoted to detect warm areas, while the other two are in charged of detecting also cold objects that potentially can be pedestrians. All these approaches build a list of rectangular bounding boxes framing interesting areas.

A following processing is in charged of localizing the homologous bounding boxes in the other image, thus allowing an estimation of bounding boxes distance and position.

Bounding boxes featuring the same distance and a similar position are the grouped together in order to build larger bounding boxes. This result is filtered using constraints about minimum and maximum allowed pedestrian size and aspect ratio.

A further match is used to search the most evident human shape characteristic in the FIR domain, the head. An overall system flow is depicted in figure 1.

2.1 Underlying approaches

In the following the three approaches for the detection of the areas of attention, the following stereo match and the filterings are described.

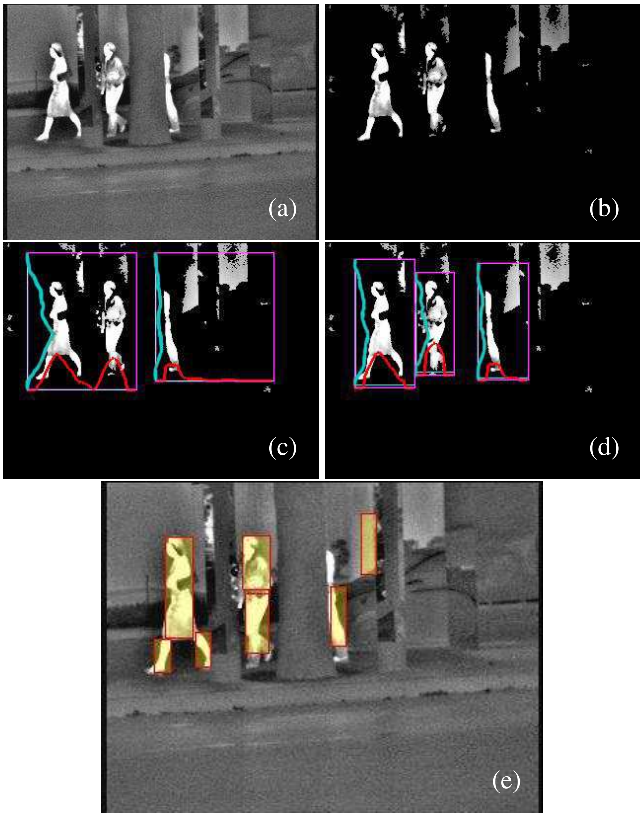

Figure 2: Warm elements detection: (*a*) original input image, (*b*) focus of attention, (*c*) intersection of column-wise and row-wise histograms produces wide rectangular bounding boxes, (*d*) a following iteration reduces their size, (*e*) final results are shown as yellow bounding boxes superimposed onto the original image.

2.1.1 Warm areas detection

The following processing is aimed at focusing the attention on areas of the input image (fig. 2.a) with a high intensity value, namely warm objects. This is obtained using a two threshold approach: initially, a high threshold is applied on the image zeroing cold or barely warm areas, thus selecting only very bright (namely warm) areas. A region growing processing is then started. Pixels contiguous to warm areas are selected if, in the original image, they feature a value higher than a low threshold.

The resulting image contains only warm contiguous areas that presents hot spots (fig. 2.b). In order to build a list of bounding boxes that contain warm areas, an iterative approach based on histograms is used.

Initially, a column-wise histogram of grey levels is computed on the resulting image. The histogram is filtered using a dynamic threshold computed as a percentage of the average value of the whole histogram. This produces a list of vertical stripes that contain warm objects. Since multiple warm areas may be vertically aligned in the image adding up their contribution to the histogram, a new row-wise histogram is computed and filtered for each stripe. Thus a number of rectangular bounding boxes framing interesting areas (see fig. 2.c) is produced.

In order to refine these bounding boxes, the column-wise and row-wise histogram procedure is iteratively applied to each rectangular box (fig. 2.d) until their size can not be longer reduced (fig. 2.e).

In this phase, small bounding boxes are thrown out, as they represents nuisance's elements.

Unfortunately, this approach fails in detecting the correct areas of interest when pedestrians are not warmer than the background or when too much warm objects are present in the scene (like in the urban environment). Therefore, two additional approaches not based on thermal features, are used to produce other areas of interest.

2.1.2 Edges detection

This approach is based on the assumption that human shape features more vertical edges than the background or than other objects. Therefore, it is based on the detection of areas that contains a high amount of vertical edges.

Initially, acquired images are filtered using a Sobel operator and an adaptive threshold to detect nearly-vertical edges (fig. 3.b). Unfortunately, other objects than pedestrians features vertical edges: buildings, cars, poles… Anyway, vertical edges of these objects are generally longer and more regular than the ones that belong to human shapes. Therefore a filtering phase devoted to the removal of regular vertical edges that are longer than a given threshold is performed. Also isolated pixels are considered as noise and removed (fig. 3.c).

In order to further enhance vertical edges and to group edges of the same pedestrians in the same cluster, a morphological expansion is performed using a 3×7 operator. In the final result (fig. 3.d), only few cluster of edges are present.

A labelling approach is used to compute connected clusters of pixels, the resulting image is analyzed and a list of bounding boxes containing connected clusters of pixels is built (fig. 3.e). Generally, in complex scenarios, a large number of small clusters is detected affecting both the effectiveness and the efficiency of the subsequent processing steps. Thus, a filtering is used to remove bounding boxes that are too small and do not contain any warm areas; only sufficiently big boxes or boxes that contain bright pixels survive to this phase (fig. 3.f).

2.1.3 *v-disparity* approach for obstacles detection

Also a *v-disparity* [9] approach is used to strengthen pedestrian detection. For each pair (left and right) of rows of the images (since optical axes are parallel, they correspond to the epipolar lines) a correlation function is computed for different offset values (disparities). The result is a new image, the v-disparity image, where the the abscissa axis plots the offset for which the correlation has been computed and the ordinates axis plots the image row number. In the v-disparity image, the brighter the pixel, the higher the correlation.

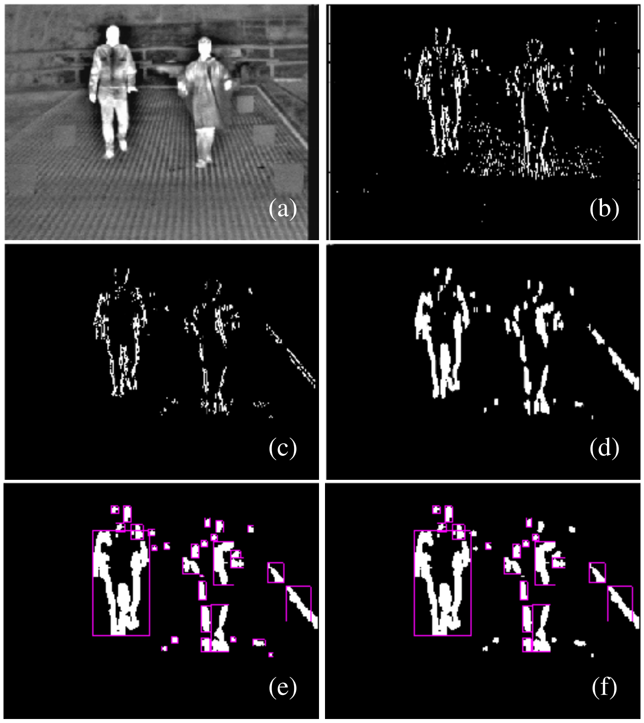

Figure 3: Edges detection: (*a*) original image, (*b*) vertical edges detected using the Sobel operator and a threshold, (*c*) edges after the removal of strong and regular edges, (*d*) expanded edges, (*e*) bounding boxes containing connected clusters of pixels, and (*f*) final resulting list of bounding boxes.

It can be noticed that objects that are present in the scene are mapped in this image as vertical bright segments. This fact allows both their detection and, thanks to 3D information, the computation of their distance as described in a previous work [10].

The result is a new image in which the pixels values encode the presence and the distance of detected obstacles (see fig. 4).

An iterative approach based on the use of histograms similar to the one described in paragraph 2.1.1 is used to build a list of bounding boxes that contain areas of attention.

Anyway, this process is fairly different form the other two previously described; in fact, in such case 3D informa-

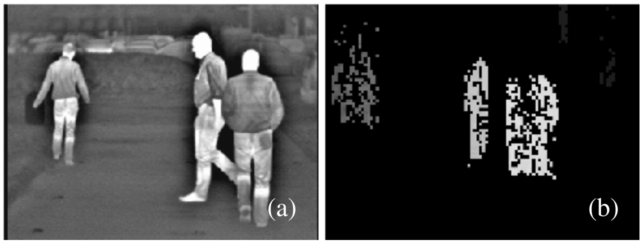

Figure 4: V-disparity approach: (*a*) original image and (*b*) detected obstacles, the darker the pixels the further the obstacle.

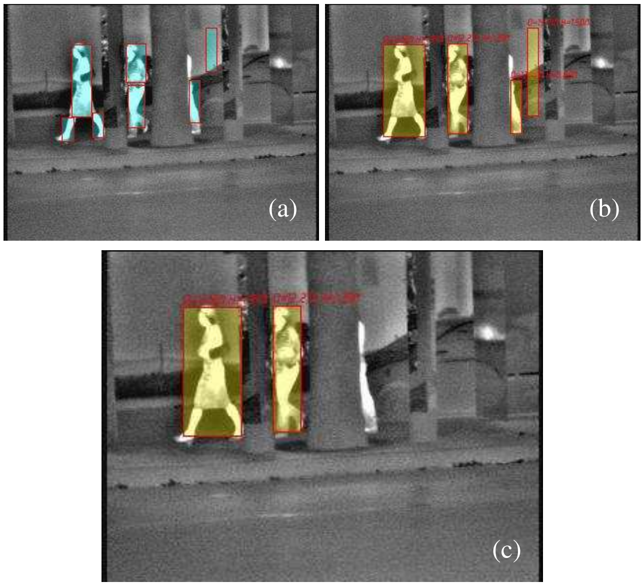

Figure 5: (*a*) Homologous bounding boxes in the left image, (*b*) bounding box bases are resized to reach the ground and elements with similar 3*D* coordinates are grouped together, (*c*) areas compatible with the presence of pedestrians.

tions about areas of attention are already available, since the detection is stereo vision-based.

Conversely, an additional step is needed for bounding boxes produced during the processings described in sections 2.1.1 and 2.1.2 in order to compute the distance and size for attention areas.

2.2 Stereo match

This phase is used to match resulting areas of attention against the other image. Once a correspondence is found, it is possible to compute size and distance of the framed object. Only areas of attention computed using the approaches described in 2.1.1 and 2.1.2 are processed during this phase, since for the ones computed using the v-disparity approach distance and size have been already computed.

Starting from the assumption of parallel optical axises, homologous areas can be localized on the same row in the two images and limited search space can be estimated thanks to the knowledge of calibration parameters.

A Pearson's correlation function is used to evaluate the quality of the match:

$$r = \frac{\sum a_{xy} b_{xy} - \frac{\sum a_{xy} \sum b_{xy}}{N}}{\sqrt{\left(\sum a_{xy}{}^2 - \frac{(\sum a_{xy})^2}{N}\right)\left(\sum b_{xy}{}^2 - \frac{(\sum b_{xy})^2}{N}\right)}} \qquad (1)$$

where N is the number of pixels in the considered bounding box, a_{xy} and b_{xy} are corresponding pixels values of the two images. The bounding box in the other image featuring the maximum value for the correlation is selected as the best match, and the algorithm considers only the best matches whose correlation values are higher than a given threshold (fig. 5.a).

A triangulation technique is used to estimate the distance between each object and the vision system.

2.3 Bounding boxes resize

The knowledge of the distances of bounding boxes allows a refinement of the boxes size. In fact, assuming a flat road in front of the vision system, it is possible to compute the point of contact between each object framed by a bounding box and the ground. Thus the bottoms of bounding boxes are stretched till they reach the ground [2].

2.4 Bounding Boxes merge and filtering

The knowledge of 3D informations allow both to merge and preliminarily filter the areas of attention.

Bounding boxes located near the same position in the 3D world are, in fact, all grouped into a single and bigger bounding box (see fig. 5.b). In addition, a number of filters have been devised to get rid of false positives. Too small or huge bounding boxes that can not contain human shapes are removed; pedestrians are expected to stand on foot, thus also bounding boxes much more large than tall are discarded.

Distance and height of each potential pedestrian are evaluated as well: too close bounding boxes that can not contain a whole human shape or too far areas that do not allow a sufficiently reliable analysis are discarded as well.

2.5 Search for pedestrians head

For each potential pedestrian the presence of the most evident feature of an human shape in the infrared domain –the head– is searched for. The position of the head is not affected by pedestrian's pose, being always in the upper part of the bounding box and it is, often, warmer than the body.

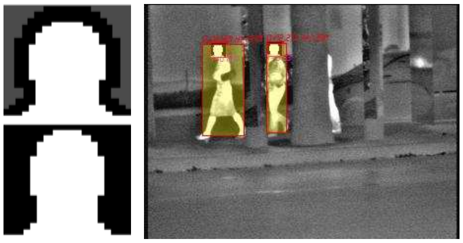

Figure 6: Head search: the two models used for head matching and best matches superimposed on areas of attention.

Two different models of a head are used to perform pattern matching operations.

The first model encodes thermal characteristics of a head warmer than the background, namely a binary mask showing a white head on a black background (fig. 6). For easing the match, also the areas of attention are binarized as well using an adaptive threshold. The model is scaled according to the bounding boxes size and assuming that a head measures nearly $1/6$ of human shape height and the match is performed against an area centered around the top of the bounding box using equation 1. The highest value obtained (P_w) is considered as match quality.

Unfortunately, not always the head is warmer than the background. Environmental conditions, hats, helmets, or hair may mask heat radiation. In order to cope with this problem, an additional head model is used. This model encodes the head shape (fig. 6) and is used to perform another match in the top area of each bounding box. In this case, the areas of attention are not binarized; for each position of the model, the two average values of pixels that correspond to the internal (white) or external (black) part of the model are computed and a the quality of the match computed as the absolute value of the difference beetween these two averages. A higher difference is obtained in correspondence to object that feature a shape similar to the model's one. The matching quality (P_s) is computed as the highest of such differences amongst the ones computed for the different model positions.

The final match parameter is computed as

$$P_m = 1 - ((1 - P_w) \times (1 - P_s)).$$

The portion of the bounding box that produces the best match value is recognized as the head of a potential pedestrian while boxes featuring a too bad match are discarded as not containing a human shape (fig. 6).

3 Results

The developed system has been tested in different situations using an experimental vehicle equipped with two Raytheon 300D 7–14μ infrared camera.

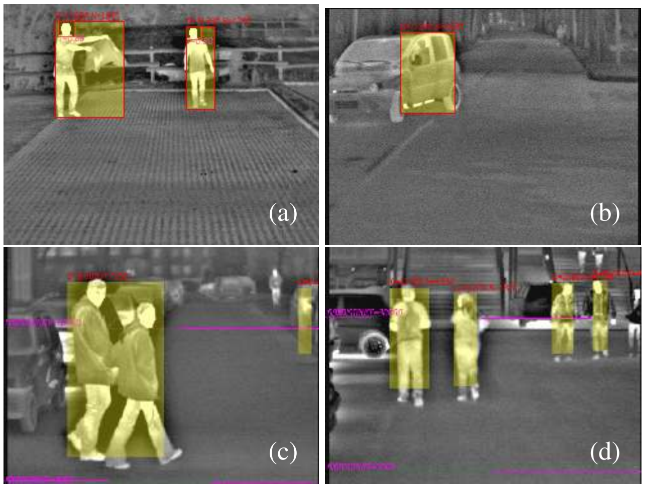

Figure 9: Detection problems: (a) shows a pedestrian with unusual positions its aspect ratio is similar with other possible objects in different situations, like in (b), (c) shows a group of pedestrians detected as a single one, and (d) present a case where an error in the calibration affects the distance computation and also the box refinement.

Fig. 7 shows few results: detected human shapes are evidenced using a superimposed yellow box, distance (D) and height (H) are indicated on the top of each box. Thanks to the triangulation information, the system has proven to be able to detect pedestrians even if the are partly overlapped each other (fig. 8.a). In addition, the use of three different approaches for the detection allows to detect pedestrians in complex scenarios or even when they are not warmer than the background.

Moreover, even though the model used for head detection is quite simple, the head of each pedestrian is properly localized. The model represents a frontal head shape but it is useful also for side shots (fig. 8). In fig. 8 pedestrian in foreground is not detected, even if visible, because its height is too small for an human shape and the corresponding bounding box has no valid aspect-ratio, so the filtering phase discards it.

Main problems to check are about aspect ratio. Sometimes aspect ratio is not a good evaluation criterion for filtering results. Figure 9 shows that a pedestrian with open arms produces a bounding box that is compatible with other possible objects in the scene, like cars. This problem can appear in different situation causing the system to get false positive results despite of the head pattern matching filtering.

Another problem concerns groups of pedestrians, if pedestrians are very close each others and at the same distance from the vision system, they are often detected as a single pedestrian (fig. 9.c). Another problem is due to the precision of calibration: any deviation affects distance and

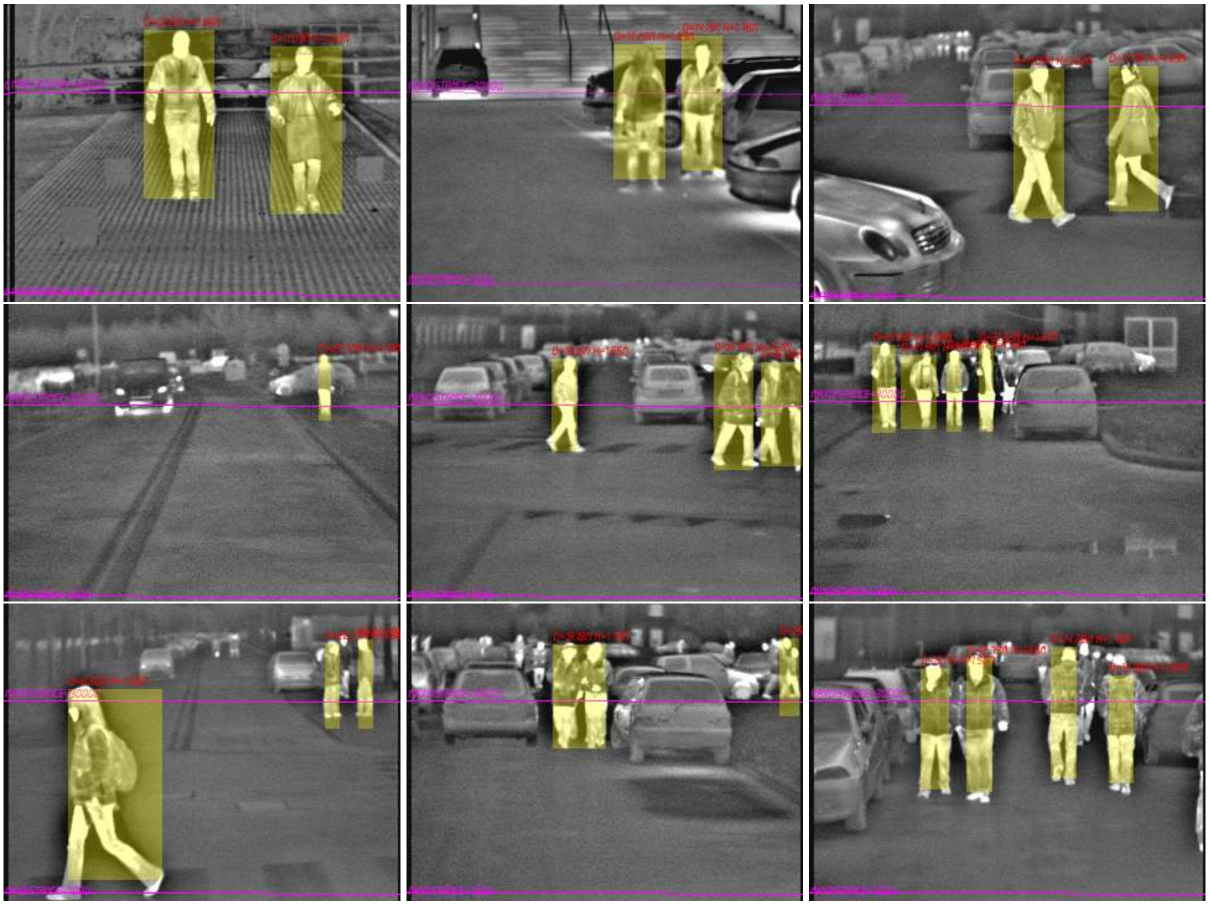

Figure 7: Pedestrian detection results: detected pedestrians are showed using a superimposed yellow box. Each box also shows the distance and height of detected object. The two magenta lines show the minimum and maximum distances used for the detection.

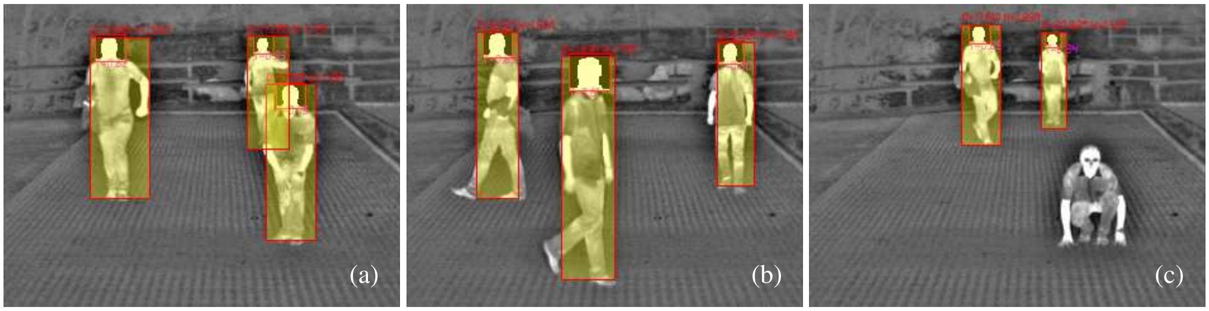

Figure 8: Results: (*a*) the system is able to detect pedestrian even if partially occluded, (*b*) the head model is appropriate whether for frontal or side shots, and (*c*) pedestrian in foreground is not properly detected due to height and aspect ratio constraints.

size computation for detected objects and subsequent steps, like the refinement of bounding boxes till ground (fig. 9.d).

Other detection failures are due to occlusions, but this problem is observed only in few frames, thus tracking could be used to cope with these particular cases.

Another source of misdetections is due to the head match filter. In fact, even if it has been improved with respect to the one discussed in [2], it is still based on thermal characteristics only and fails when the head is colder than the background. A new head filter based on shape characteristics only is currently under development.

The system has been tested using a tool for performance evaluation [1], a ROC curve is shown in fig. 10 varying the correlation threshold used for the detection. The first curves (fig. 10.a) has been obtained running the system enabling warm, edges and v-disparity approaches. It can be noticed that in this condition the system is able to correctly detect more than 80% of pedestrians in the scene maintaining a low value for false detections.. The second ROC curve (fig. 10.b) has been obtained using warm area detection only, and thus shows how edge and v-disparity approaches improve the overall result while barely affecting false detections.

Tests have been performed using a Pentium IV processor at 2.80GHz equipped PC with 512 kBytes of cache memory and 1 GByte of RAM. Since processing time strongly depends on how much the scene is complex, both urban and extra-urbane sequences have been used to evaluate temporal performance; the average execution time on the reference architecture is 84 ms; that means that the system is able to process nearly 12 frames per second.

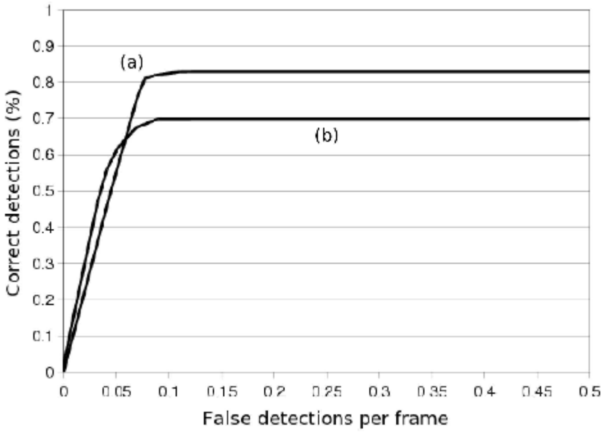

Figure 10: ROC curves of the results with (*a*) all underlying approaches activated or (*b*) only warm areas detection.

4 Conclusions

In this paper a stereo vision-based algorithm aimed at the detection of pedestrian in infrared images has been discussed. It has been tested in urban and extraurban environments using an experimental vehicle equipped with two infrared camera that work in the 7–14μ spectrum.

The algorithm is based on three different approaches: the detection of warm areas, the detection of vertical edges and a v-disparity computation. Distance estimation, size, aspect ratio, and a head presence are used to select pedestrians. Neither temporal correlation, nor motion cues are used for the processing.

Experimental results demonstrated that the approach is promising. The presence of two additional approaches not based on the detection of thermal characteristics permits to increase the detection ratio with respect to the previous approach [2] and to detect pedestrians even if they are not warmer than the background.

Correct detection percentage is high with a very low number of false detection per frame, and the system has proven to work also when pedestrians are partly occluded.

In urban situations, noise produced by the presence of buildings, cars, signals and other objects could increase false detections. In order to enhance the robustness and reliability of the discussed system, a tracking algorithm abd an improved head filter are currently under development.

References

[1] M. Bertozzi, A. Broggi, P. Grisleri, A. Tibaldi, and M. D. Rose. A Tool for Vision based Pedestrian Detection Performance Evaluation. In *Procs. IEEE Intelligent Vehicles Symposium 2004*, pages 784–789, Parma, Italy, June 2004.

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

[3] A. Broggi, M. Bertozzi, , R. Chapuis, F. C. A. Fascioli, and A. Tibaldi. Pedestrian Localization and Tracking System with Kalman Filtering. In *Procs. IEEE Intelligent Vehicles Symposium 2004*, pages 584–589, Parma, Italy, June 2004.

[4] Y. Fang, K. Yamada, Y. Ninomiya, B. Horn, and I. Masaki. Comparison between Infrared-image-based and Visible-image-based Approaches for Pedestrian Detection. In *Procs. IEEE Intelligent Vehicles Symposium 2003*, pages 505–510, Columbus, USA, June 2003.

[5] K. Fujimoto, H. Muro, N. Shimomura, T. Oki, Y. K. K. Maeda, and M. Hagino. A Study on Pedestrian Detection Technology using Stereo Images. *JSAE Review*, 23(3):383–385, Aug. 2002.

[6] D. M. Gavrila and J. Geibel. Shape-Based Pedestrian Detection and Tracking. In *Procs. IEEE Intelligent Vehicles Symposium 2002*, Paris, France, June 2002.

[7] Y. L. Guilloux and J. Lonnoy. PAROTO Project: The Benefit of Infrared Imagery for Obstacle Avoidance. In *Procs. IEEE Intelligent Vehicles Symposium 2002*, Paris, France, June 2002.

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

[9] R. Labayrade, D. Aubert, and J.-P. Tarel. Real Time Obstacle Detection in Stereo Vision on non Flat Road Geometry through "V-Disparity" Representation. In *Procs. IEEE Intelligent Vehicles Symposium 2002*, Paris, France, June 2002.

[10] U. Ozguner, K. A. Redmill, and A. Broggi. Team TerraMax and the DARPA Grand Challenge: A General Overview. In *Procs. IEEE Intelligent Vehicles Symposium 2004*, pages 232–237, Parma, Italy, June 2004.

[11] V. Philomin, R. Duraiswami, and L. Davis. Pedestrian Tracking from a Moving Vehicle. In *Procs. IEEE Intelligent Vehicles Symposium 2000*, pages 350–355, Detroit, USA, Oct. 2000.

[12] Y. Song, X. Feng, and P. Perona. Towards detection of humans. In *Procs. Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 810–817, South Carolina, USA, June 2000.

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

[14] L. Zhao and C. Thorpe. Stereo and neural network-based pedestrian detection. *IEEE Trans. on Intelligent Transportation Systems*, 1(3):148–154, Sept. 2000.