

Stereo-based Preprocessing for Human Shape Localization in Unstructured Environments

A. Broggi, A. Fascioli, I. Fedriga, and A. Tibaldi

Dipartimento di Ingegneria dell'Informazione
Università di Parma
Parma, I-43100, Italy
{broggi,fascal,fedriga,tibaldi}@ce.unipr.it

M. Del Rose

Electronic Research
Vetronics Research Center, U.S.Army TACOM
Warren, MI, U.S.A.
DelRoseM@tacom.army.mil

Abstract— This paper describes the research activities for the localization of human shapes using visual information carried on at the University of Parma, Italy, in the frame of a common project with the TACOM Department of U. S. Army.

The paper proposes the application of a stereoscopic technique as a preprocessing for the localization of humans in generic unstructured environments. Each row of the left image is matched with the epipolar row of the right image. This creates a map of each object in the scene as well as the slope of the road. Preliminary results have proved to be promising.

Keywords— Pedestrian Detection, Stereo Vision, Symmetry, Obstacle Detection

I. INTRODUCTION

AUTONOMOUS navigation will be a vital part in the near future for the U. S. Army. The Vetronics Technology Area, a division within the Army's TACOM Research, Development, and Engineering Center (TARDEC), and the University of Parma are working towards detect human shapes from a moving vehicle. Programs within Vetronics have a need for autonomous navigation, the Robotic Follow program [1], and semi-autonomous navigation, the Crew integration and Automation Test bed program [2]. Human shapes detection is a vital piece to make these programs within the Vetronics Technology Area succeed.

The Robotic Follower is a robotic vehicle that is used to follow behind a person or another vehicle to carry supplies to and from areas [1]. The path the robotic follower takes is based on electronic breadcrumbs that are left behind by the lead person or vehicle. The greater the distance between the leader and the follower, the more chance of people interrupting the breadcrumb trail. The Robotic Follower needs the ability to detect people in and around it so that it can take the necessary precautions to avoid them.

The Crew integration and Automation Test bed program is designed to incorporate several driver aided packages [2], one of which is detecting humans. As the driver traverses through an area, the system will detect people, and highlight them, so that the driver is more aware of their presence.

However, the need for automatic human detection goes beyond the Vetronics Technology Area. Within the Army, and throughout the commercial community, the need for detecting people is great. Driver awareness systems, security systems, traffic/pedestrian control systems, and automatic switching sys-

This work was supported by the European Research Office of the U. S. Army under contract number N62558-02-M-6017.

tems are just a few areas that would benefit greatly from this technology.

There are many approaches to pedestrian detection. Some use learning machines like neural networks [3] or support vector machines [4], some use motion to detect pedestrians [5, 6]. Throughout this study, motion was not used. The authors felt that approaching the problem of solving the cases for individual frames will prove to be more beneficial than using continuous frames. Tracking can be added later to reduce the number of false positives. The specific application of the method shown in this paper is unique to the stereo vision community. Each row of the left image is matched with the corresponding row of the right image. This creates a map of each object in the scene as well as the slope of the road. Both information can be used in the human shape localization algorithm presented in [7]. Preliminary results have proved to be promising.

This paper is organized as follows: section II introduces the vision-based system for detecting pedestrians in road environments developed in the last years by the University of Parma in collaboration with TACOM. Section III presents a stereo based technique for the extraction of features of interest. The results of this approach are shown in section IV, while its application and advantages are discussed in section V.

II. A STEREO-BASED APPROACH IN STRUCTURED ENVIRONMENTS

IN the last years the University of Parma and the TACOM Department of U. S. Army developed a vision-based system for detecting pedestrians in road environments [8, 7]. The system is aimed at the localization of pedestrians in various poses, positions and clothing, and is not limited to moving people.

Attentive vision techniques relying on the search for specific characteristics of pedestrians, such as vertical symmetry and strong presence of edges, are used to select interesting regions likely to contain pedestrians. More precisely, the acquired image is scanned and symmetries and edges are extracted; since a human shape is characterized by a strong vertical symmetry, symmetrical areas with a specific aspect ratio identify possible candidates. Thanks to some a-priori knowledge on the environment (the slope is known since the road is assumed flat), size and perspective constraints are also adopted to ease and speed up the search.

Specific filters are then used to remove evident detection errors and false positives.

Subsequently, the remaining candidate areas are validated verifying the actual presence of pedestrians by means of shape-based techniques. A method based on the application of autonomous agents has been investigated [7], and other approaches are under study.

This system, completely based on monocular techniques, has subsequently been enhanced thanks to a stereo-based refinement. In fact, some errors may arise in the first phase due to an incorrect localization of candidates. In other words, a human body may present a sufficiently high symmetry to be detected, but the detected area may not be precise. This generally happens to the legs, which can be in different positions. In these cases, a bounding box enclosing the human body is drawn around the detected shape, but it may cut out a part of the body –generally the legs.–

An incorrect localization of the bounding box may be critical for the following shape detection process aimed at its validation. Moreover, this error affects distance estimation in monocular images. The stereo refinement is targeted to fix this problem in the assumption of a flat road.

First, the left image is searched for symmetries, bounding boxes corresponding to candidates are generated, and a set of filters are applied to remove obvious errors in the detection. Then, for each surviving bounding box the right image is searched for areas which exhibit a content similar to the one included in the bounding box (a correlation measure is performed). Once the correspondence between the bounding box located in the left image and its counterpart in the right image has been found, stereoscopy can be used to determine the distance to the vision system. This step requires the correct calibration of cameras parameters and orientations.

Once the correct distance estimation for each bounding box has been provided to the system by means of stereoscopy, a refinement of the bounding box base can take place, based on calibration and perspective constraints. More precisely, the knowledge of the camera orientation with respect to the ground and the road slope can provide information about the position of the point of contact of the human shape with the ground. This knowledge is used to stretch the bottom of the bounding box till it reaches the ground and frames the entire shape of the pedestrian, thus easing the following shape-based validation. Figure 1 shows the result of this stereo refinement.

III. A STEREO TECHNIQUE FOR FEATURE EXTRACTION

THE current research addresses the problem of human shape localization in generic environments, including urban, country, and desolate.

The previously discussed approach could be easily generalized to any scenarios (including non flat ones) removing the assumption on the knowledge of the road slope. In this case, however, size and perspective constraints are to be dropped, and an exhaustive search has to be performed in the candidates generation phase. This entails both a higher computational complexity and a more complex selection of the interesting areas since a high number of candidates must be considered and compared.

Moreover, the stereo refinement of the bounding boxes, as defined previously, is not possible if the scene slope is unknown, and possible errors in the bounding box localization are to be

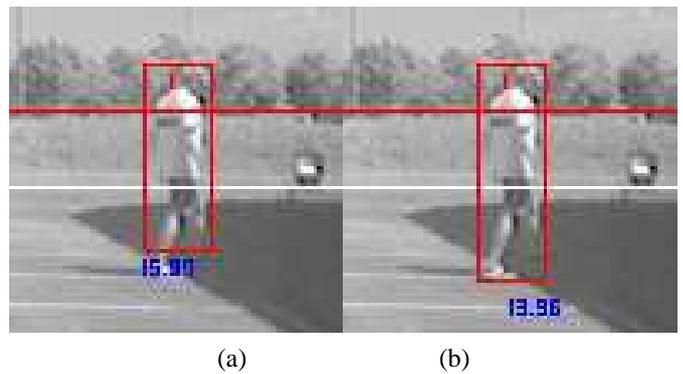


Fig. 1 Distance refinement: (a) result before refinement: a potential pedestrian is detected but the bounding box cuts the legs thus affecting the distance estimation; (b) result after stereo refinement: the bounding box has been stretched till the ground; the distance estimation is now correct.

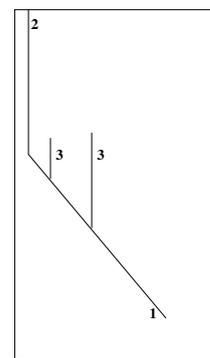


Fig. 2 Different components of correlation: (1) ground slope, (2) background, and (3) obstacles.

tackled in the following shape-based validation step.

Following these considerations, a new stereoscopic approach has been developed to deal with generic environments where the scene slope is unknown, featuring low computational complexity. This method is based on a row-wise comparison of the two stereo images, assuming the two optical axes lie on the same plane and both cameras have a null roll angle.

This approach has been tested on both synthetic and real images, see figure 3. The left and right images (figures 3.a and 3.b) are processed with the following steps: an edge extraction, followed by a binarization and a morphological horizontal expansion are performed. The results are pixel-ORed with the original images. In this way, the original grey-level values are only preserved in correspondence to areas with a relevant information content (i.e. edge points). Figures 3.c show the result of the processing of figures 3.a (left images); the same processing is applied to the right images. The resulting images will be referred to in the following as *feature images*.

For each line the correlation between the left and right epipolar lines of the feature images is computed for different offsets. Figures 3.d show the value of the correlation of each image line for different offsets: these *correlation images* display the offset on the horizontal axis and the image line number on the vertical axis, encoding high correlation values with bright pixels.

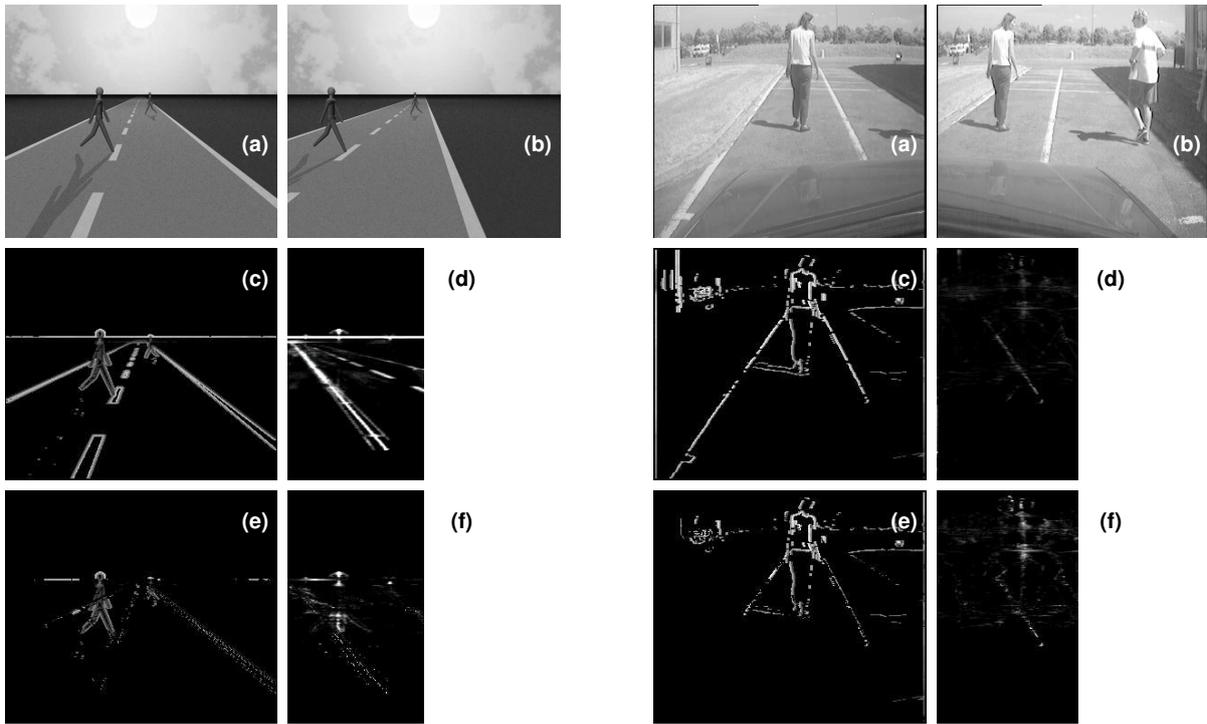


Fig. 3 Obstacle detection in a synthetic and a real situation: (a) left and (b) right images, (c) relevant features computed for the left images, (d) line-wise correlation values between left and right features images for different offsets, (e) left features image after the removal of background, (f) line-wise correlation values computed after the removal of background.

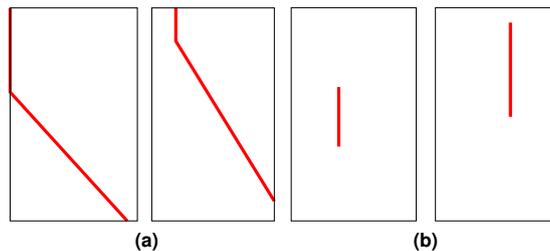


Fig. 4 Reconstruction of correlation components extracted from figures 3.d: (a) road and background components, (b) components given by the closest pedestrians.

Perspective considerations allow to discriminate different components (modeled in figure 2) in this correlation image:

1. a slanted line encodes the ground slope (in this cases, the road),
2. a vertical line on the top left of the image encodes the background (above the horizon), and
3. other vertical segments originating upward from the slanted line represent potential obstacles (in this case, pedestrians).

These components are hardly discernible as shown in figures 3.d, because they mask each other; moreover, noise affects the correlation measure.

The following procedure is aimed at extracting them one by one.

In fact, the strongest component of the correlation encodes the longitudinal slope of the scene, provided that the transversal slope of the scene is neglectable. For example, in case of a flat scene without obstacles the offset yielding the maximum

correlation decreases with the distance from the vision system according to a known function (component 1), and becomes constant in correspondence to the horizon and upper (component 2) [9]. This behavior is due to the fact that the difference of displacement in left and right images for 3D points lying close to the camera is larger than for 3D points lying far away from the vision system. On the other hand, 3D points at infinite distance are imaged in the same position in the left and right images when the two optical axes are parallel, or at a constant offset in case the axes are convergent or divergent.

Components 1 and 2 are evident when the ground surface and background present an appreciable texture. When present, they prevail and partially mask the other obstacles' components. Each obstacle contributes to a vertical segment in the correlation image correspondent to a constant offset. However, components 1 and 2 are stronger than the obstacle's one. This effect can be exploited to identify and remove the ground and back-

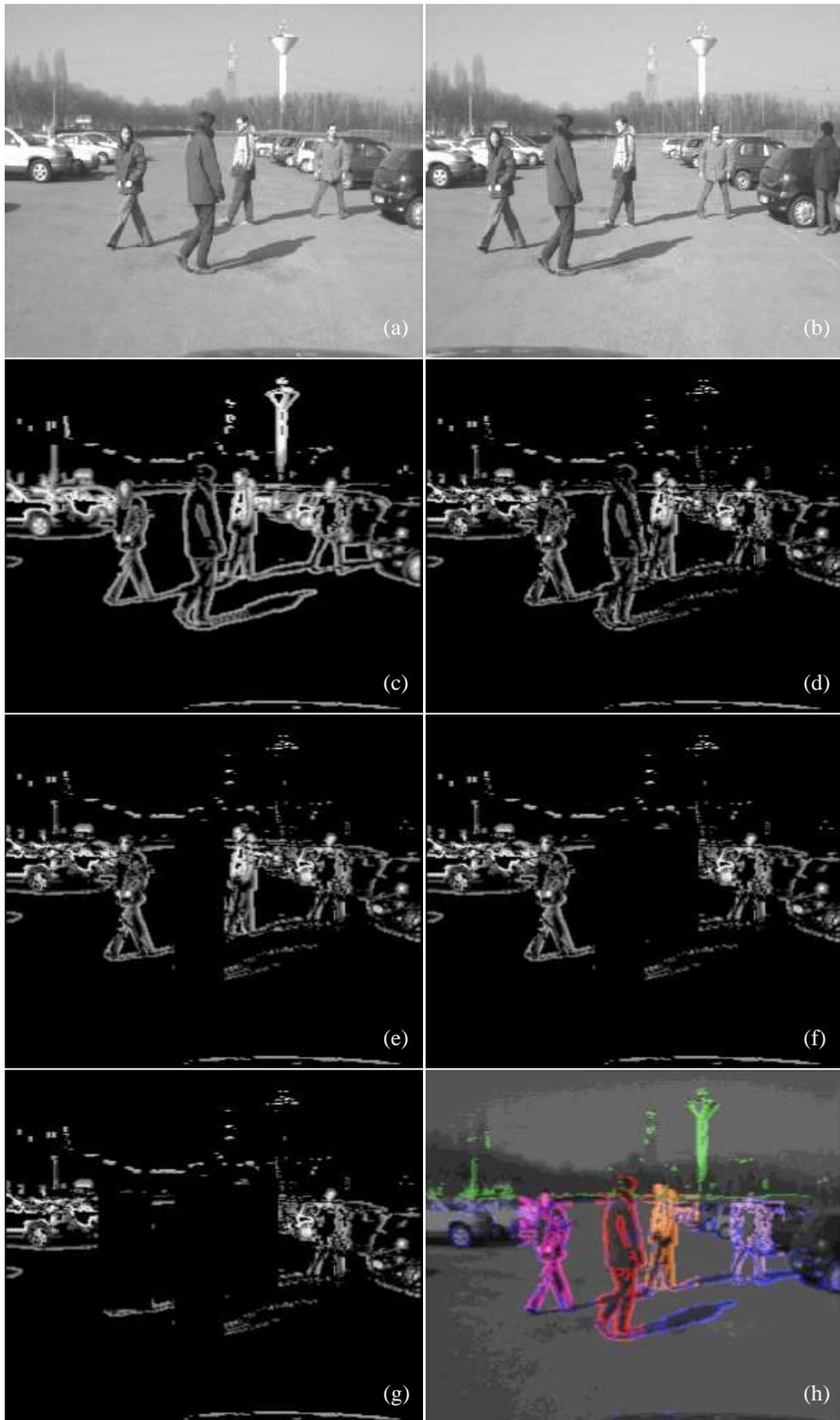


Fig. 5 (a) and (b) Left and right images, (c) relevant features computed for the left images, (d) removal of ground texture and background, (e) removal of first object, (f) removal of second object, (g) removal of third object, (h) the clusters of features labeled with different colors; the ground features are shown in blue, while background objects are highlighted in green.

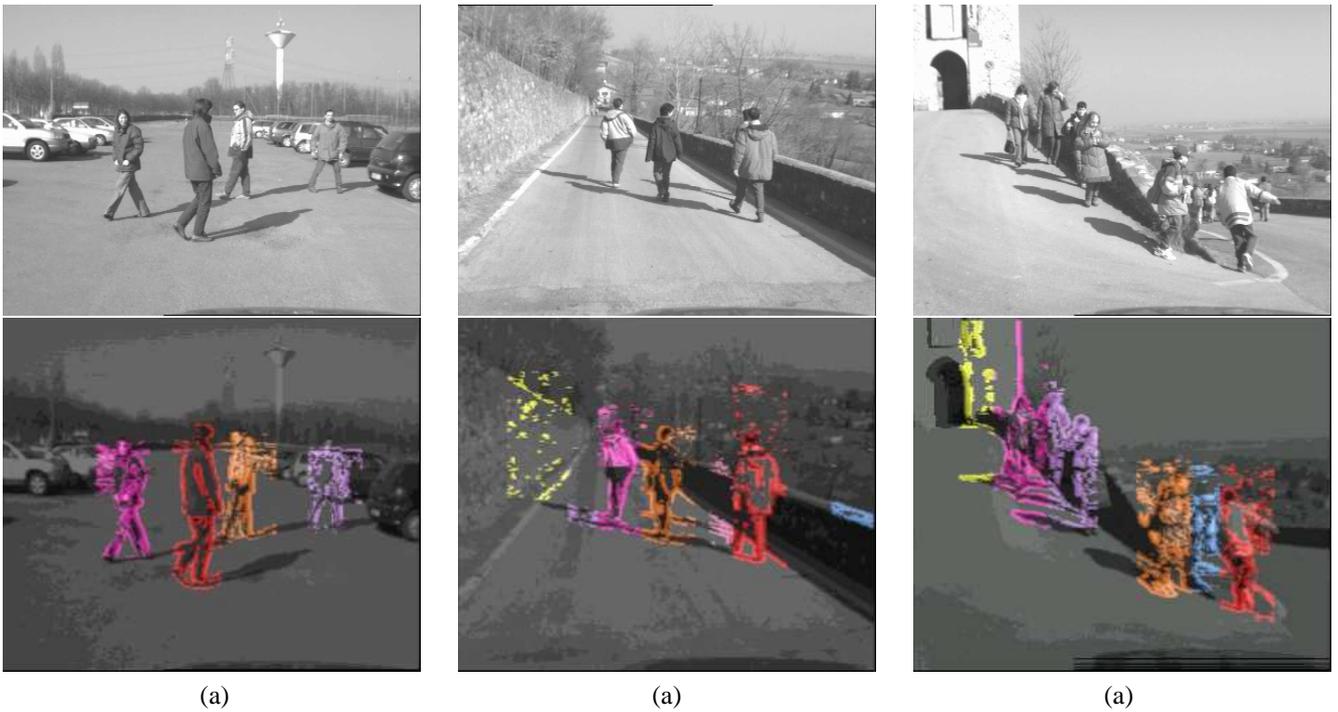


Fig. 6 Results in different situations.

ground features leaving only the features belonging to the obstacles. The correlation image is thus analyzed to detect its components. More specifically, the slanted line corresponding to the road slope and the vertical line representing the background offset are first extracted by using a Hough transform on the correlation image. Figures 4.a show a reconstruction of components 1 and 2 extracted from figures 3.d in the synthetic and real case.

The set of offsets encoded in this polyline are then applied to match the epipolar lines of the left and right feature images. This comparison is used to remove matching features (i. e. ground texture and background objects), thus leaving non-matching areas (i. e. foreground obstacles). Figures 3.e show the left features images after the removal of ground and background features; it can be noticed that obstacles are more evident since many disturbing features have been filtered out.

The computation of the correlation image is then repeated, starting from the feature images with ground and background removed (see figures 3.f). Now, the obstacle components prevail in the correlation image and can be extracted thanks to a vertical histogram. Figures 4.b show a synthetic reconstruction of the correlation values for the closest obstacles.

Once its offset is known, the area of each obstacle can be further analyzed to derive the cluster of features belonging to it. The right feature image is shifted with the offset corresponding to the obstacle and compared to the left, and their matching features are examined. A vertical histogram allows the identification of the position of the lateral borders of the obstacle. A horizontal histogram computed in the vertical stripe the object belongs to gives hints on the bottom and top limits of the object.

In presence of multiple objects lying at different distances from the vision system, the localization of individual objects is simplified if the previously segmented objects are in turn elim-

inated from the feature image and the correlation function is every time recomputed. In this manner, the strong contribution to the correlation given by an evident object does not mask weaker contributions given by other objects, and objects can be extracted in subsequent iterations of the processing. At each stage the features belonging to a different obstacle can be clustered and labeled. Figures 5.c-g show how the features belonging to the ground and background and three different obstacles are identified and removed at subsequent steps of the processing. In figure 5.h the detected clusters of features have been labeled with different colors.

IV. RESULTS

Figures 6 and 7 show examples of extraction of obstacles' features from stereo images in unstructured environments. The original left image is displayed together with a copy with obstacles' edges highlighted with different colors.

Figure 6 presents three examples of correct detection, detailed in the following. Figure 6.a shows a parking lot with four pedestrians and vehicles on the sides: edges of pedestrians are correctly localized. Figure 6.b displays descending ramp: four objects are localized (three pedestrians and a short wall on the right), the wall on the left is also detected despite its weak texture. Figure 6.c side view of a steep descending ramp with a group of children: humans are localized. Note that the high transversal slope causes an incorrect detection of the top left group of children (shadows are misinterpreted as belonging to obstacles).

Figure 7 shows some problems of the current version of the algorithm. In particular, figure 7.a presents an off-road country environment: the very weak texture of the ground does not allow the determination of its slope and the resulting cluster of

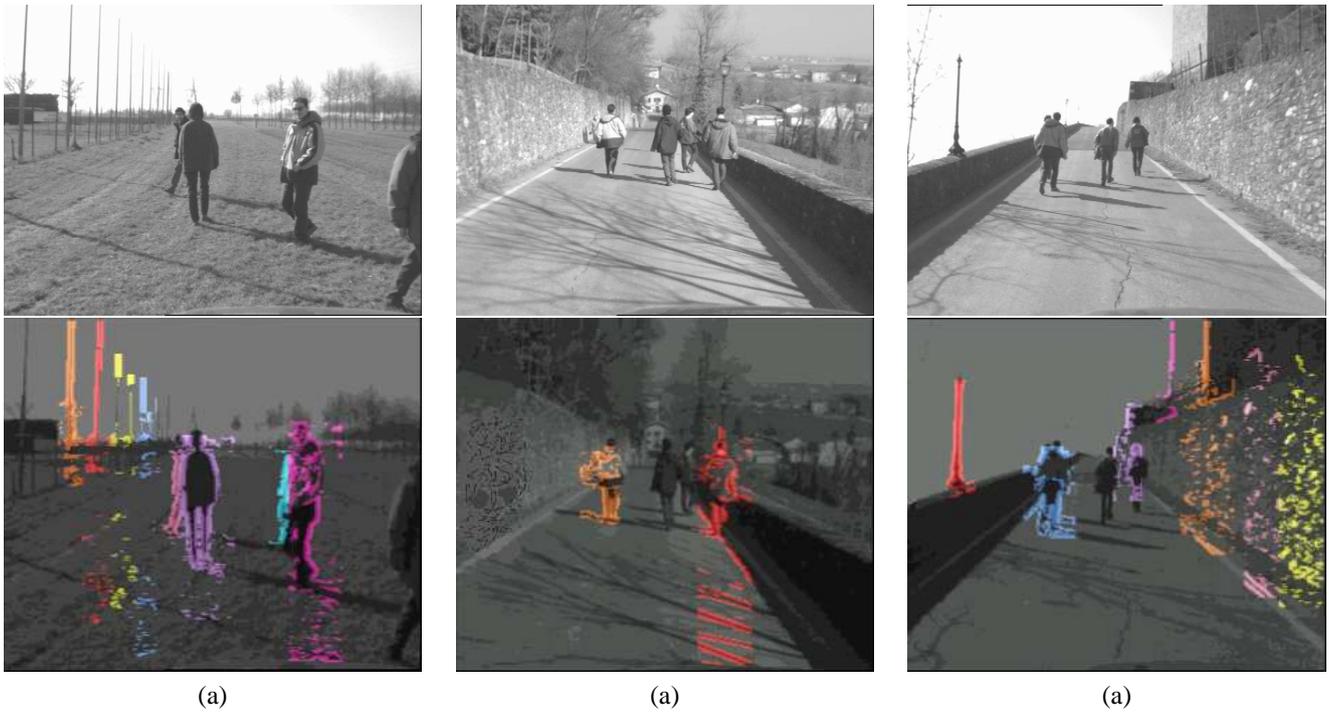


Fig. 7 Situations in which the features extraction experience problems.

pixels include also features of the ground under the obstacle. Figure 7.b shows a descending ramp with trees' shadows: one pedestrian is not detected due to low contrast and the problems in the detection of the ground slope generate a too large cluster for the pedestrian on the right. Figure 7.c refers to uphill driving: one pedestrian is not detected due to low contrast and the long wall on the right is correctly detected as an obstacle but sliced in three parts due to the large range of distances (offsets) covered.

V. DISCUSSION

THE stereo technique discussed in section III provides clusters of features that can be fed into the original monocular processing described in section II aimed at distinguishing human shapes from other obstacles. The inclusion of this preprocessing allows to limit the computation of symmetries to the detected area of interest only, improving computational time.

Moreover, the preprocessing ability to determine the ground slope permits both the application of perspective considerations and the correct detection of the objects' point of contact with the ground.

This approach has the advantage to adapt the original method to generic scenarios. Furthermore, since the scene slope can be obtained from the stereo row-wise correlation, different approaches could be developed for flat and non-flat scenarios.

REFERENCES

- [1] B. Brendle and J. Jaczkowski, "Robotic Follower: Near-Term Autonomy for Future Combat Systems," in *Procs. SPIE - AeroSense Conference 2002*, vol. 4715, (Orlando, FL), Apr. 2002.
- [2] B. Brendle, "Cockpit Development in the Crew Integration and Automation Testbed advanced Technology Development Program," in *Procs. SPIE - AeroSense Conference 2003*, vol. 5080, (Orlando, FL), Apr. 2003.

- [3] L. Zhao and C. Thorpe, "Stereo and neural network-based pedestrian detection," *IEEE Trans. on Intelligent Transportation Systems*, vol. 1, pp. 148–154, Sept. 2000.
- [4] S. Kang, H. Byun, and S.-W. Lee, "Real-Time Pedestrian Detection Using Support Vector Machines," *Lecture Notes in Computer Science*, vol. 2388, p. 268, Feb. 2002.
- [5] V. Philomin, R. Duraiswami, and L. Davis, "Pedestrian Tracking from a Moving Vehicle," in *Procs. IEEE Intelligent Vehicles Symposium 2000*, (Detroit, USA), pp. 350–355, Oct. 2000.
- [6] Y. Song, X. Feng, and P. Perona, "Towards detection of humans," in *Procs. Conf. on Computer Vision and Pattern Recognition*, vol. 1, (South Carolina, USA), pp. 810–817, June 2000.
- [7] M. Bertozzi, A. Broggi, A. Fascioli, and P. Lombardi, "Vision-based Pedestrian Detection: will Ants Help?," in *Procs. IEEE Intelligent Vehicles Symposium 2002*, (Paris, France), June 2002.
- [8] M. Bertozzi, A. Broggi, A. Fascioli, and M. Sechi, "Shape-based Pedestrian Detection," in *Procs. IEEE Intelligent Vehicles Symposium 2000*, (Detroit, USA), pp. 215–220, Oct. 2000.
- [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.