Stereo based ROIs Generation for Detecting Pedestrians in Close Proximity

Meiqing Wu, Siew-Kei Lam, *Member, IEEE* and Thambipillai Srikanthan, *Senior Member, IEEE*

***Abstract*— Region of Interests (ROIs) generation plays a critical role in pedestrian detection systems. The challenge lies in generating as few ROIs as possible at low computational complexity, while ensuring none of the pedestrians in the scene are omitted. However, existing ROIs generation methods either result in a large number of irrelevant ROIs or are compute-intensive. In addition, distinguishing pedestrians who are in close proximity is still a big challenge. In this paper, we propose an efficient stereo based ROIs generation method that is based on a two-level incremental segmentation strategy. An adaptive strategy is first employed to identify a minimal set of clusters from the u-disparity image. The initial clusters are then further refined to distinguish pedestrians in close proximity. Experimental results based on a challenging benchmark show that the proposed algorithm outperforms two state-of-art baseline algorithms by being able to distinguish pedestrians in close proximity with a small number of ROIs.**

I. Introduction

Vision based pedestrian detection is a fundamental task in intelligent vehicle applications that aim at improving road safety. The general steps for this task are described as follows: First, a set of ROIs is generated after some pre-processing. A ROI is an image sub-region that is likely to contain a pedestrian. For each ROI, some representative features of a pedestrian are extracted and fed into a trained classifier, which will indicate whether the current ROI contains pedestrian or not. Finally, pedestrians are temporally tracked in order to monitor their trajectories and predict their behaviors. Despite its significance to the performance of detecting pedestrians, ROIs generation has received less attention in the literature compared to the other steps (i.e. pedestrian classification and tracking) [1, 2].

ROIs generation plays a key role in improving the performance of pedestrian detection by filtering out regions that do not contain pedestrians as early as possible using simple strategies. On one hand, as few ROIs as possible are desired to reduce the burden of the subsequent compute-intensive classification step. On the other hand, it is essential for the generated ROIs to encompass all the regions containing pedestrians, as failure to do so can significantly hamper the detection rate. The challenge in meeting these two conflicting goals lies in the fact that the urban driving environment is highly cluttered and dynamic. Pedestrians' appearances vary significantly as they can change their pose during walking and adorn attire of different style or color. The pedestrians in a scene also do not have uniform scales due to differences in individual heights and the distances from camera. In addition, a pedestrian may be positioned close to other objects and gets occluded. Therefore, generating as few ROIs as possible without omitting any pedestrians is still a highly challenging problem, especially in crowded scenes [1, 2]. Furthermore, the strict form factor and real-time requirements of intelligent vehicle systems necessitate low complexity algorithms for ROIs generation.

In this paper, a simple and efficient stereo based ROIs generation method is proposed. Preprocessing is first performed to obtain the Space of Interest (SOI), which reduces the search space for segmentation. The novelty of the proposed method lies in a two-level incremental segmentation on the u-disparity image. In the first level of segmentation, an adaptive connected component labeling technique is used to rapidly determine a minimal set of clusters from the coarse grained u-disparity image. This ensures that objects are not partitioned into multiple parts due to the limited fidelity of disparity map. A second level segmentation is then applied to the fine grained u-disparity image to further segregate the clusters based on a more stringent connectivity in order to distinguish pedestrians who are in close proximity. Finally, the ROIs are generated from the resulting clusters. This approach enables pedestrians in close proximity to be distinguished with a small number of ROIs.

This paper is structured as follows: Section II reviews the existing works in ROIs generation; Section III describes the system setup. The proposed ROIs generation method is presented in Section IV. In Section V, we present the experimental results to evaluate the performance of the proposed algorithm with two baseline algorithms. Finally, we conclude the paper in Section VI.

II. Related Works

In general, ROIs generation methods can be divided into two categories: monocular vision based and stereo vision based ROIs generation.

For monocular vision based setup, the simplest ROIs generation technique is the so called sliding window scheme, where the initial object hypotheses are generated by shifting the detection window over the whole image at various locations and scales [3, 4]. Some prior knowledge such as flat road assumption, camera pose and target object size are utilized to restrict the search space [5]. The sliding window scheme does not perform explicit image segmentation according to the scene geometry structure and will therefore result in very large number of ROIs, the majority of which correspond to non-pedestrian regions.

The main benefit of stereo vision over monocular vision is that the former can recover the depth information which is helpful for determining the scene geometry and filtering out

Meiqing Wu, Lam-Siew Kei and Thambipillai Srikanthan are with Centre for High Performance Embedded Systems, Nanyang Technological University, Singapore (e-mail: wume0007@e.ntu.edu.sg).

many irrelevant regions. In general, stereo vision based road surface detection is utilized to remove the irrelevant image regions first in order to restrict the search space.

The work in [6-8] multiplexes the disparity map into N discrete depth ranges. Each binary image that is associated with a certain depth range is scanned using sliding windows. The windows where the number of depth features exceeds a certain threshold are regarded as ROIs. This method is a variant of the sliding window scheme, and therefore leads to a large number of ROIs. The u-v disparity image based framework has been widely used for obstacle detection in the context of intelligent vehicles [9-12]. In these works, the combination of a vertical line in the v-disparity image and corresponding continuous span in the u-disparity image results in a single ROI. Overlapping ROIs whose associated disparity values are close will be merged.

Instead of working in the u-v-disparity space, some works rely on the X-Y-Z Euclidean space, which is reconstructed from u-v-disparity space. [13] utilizes the region-growing algorithm to segment the density map of 3D points in order to determine the possible object candidates. However, the details about the region-growing method are not described in their paper. Generic compute intensive data clustering techniques are also adopted to find the pedestrian candidates. In [14, 15], pedestrians in 3D space are modeled by means of Gaussian distribution and the subtractive clustering method is utilized to find high-density regions of the scene points in 3D world space. It is worthy to note that 3D world point reconstruction itself is also a compute intensive process. In addition, the noise in the data will be further amplified during the reconstruction process.

Urban driving environment presents a highly cluttered scenario where pedestrians are often in close proximity to other objects. In such scenarios, all of the above techniques, except the sliding window scheme, will easily result in merging objects that are in close proximity. This leads to misinterpretation during classification and tracking. In this paper, a simple and efficient u-v-disparity image based ROIs generation method is proposed. The novelty of such method lies in the coarse-to-fine segmentation strategy which fully exploits the property of disparity.

III. System Setup

In this section, we will briefly describe the basic principles of stereo geometry and the u-v disparity images. These principles serve as the mathematical foundations of the proposed algorithm presented in Section IV.

Assuming a small pitch angle, the relationship between disparity d and depth Z can be derived as follows:

$$d = \frac{fb}{Z} \tag{1}$$

Where b is the stereo baseline; f is the focal length measured in pixel.

From (1), we can see there is an inverse proportion relationship between disparity d and depth Z. This implies that d is another measure for encoding distance information. The distance interval corresponding to a minor change in d is shown in (2):

$$\Delta Z_{(d,d+1)} = \frac{fb}{d} - \frac{fb}{d+1} = \frac{fb}{d*(d+1)} \tag{2}$$

Assuming $d' = k * d$, then

$$\Delta Z_{(d\prime,d\prime+1)} = \frac{fb}{d\prime} - \frac{fb}{d\prime+1} = \frac{fb}{kd*(kd+1)} \approx \frac{1}{k^2}\Delta Z_{(d,d+1)} \tag{3}$$

Equation (3) indicates that a minor change in d correspond to a minor change in Z in the near region and a large change in depth Z in the far region.

In order to cover the same distance interval as defined in (2) from d', i.e., in order to obtain (4)

$$\Delta Z_{(d,d+1)} = \Delta Z_{(d\prime,d\prime+x)} \tag{4}$$

x should be as shown in (5)

$$x = \frac{k^2 d}{d+1-k} \tag{5}$$

Disparity map is recovered by stereo matching algorithm. A state-of-art stereo matching algorithm is proposed in [16] which achieves a good balance between reconstruction accuracy and run-time performance. It is worthy to note that this algorithm can achieve sub-pixel accuracy.

The concepts of u-disparity image and v-disparity image are first proposed in [9]. For the example disparity map shown in Fig. 1 (a), the corresponding v-disparity image and u-disparity image are shown in Fig. 1 (b) and Fig. 2(a) respectively. It can be observed that the v-disparity image provides a side-view projection of the 3-D scene. The u-disparity image, on the other hand, provides a top-view projection of the scene. Up-right obstacle points with the same X and Z values will converge onto the same position in the u-disparity image, therefore producing peak regions. The proposed algorithm aims to detect these peak regions in the u-disparity image in order to build candidate pedestrian areas.

IV. Proposed Algorithm

In this section, we propose a novel stereo-based ROIs generation algorithm that is capable of generating a small set of ROIs and differentiating pedestrians in close proximity.

The proposed ROIs generation method consists of three steps: 1) Generation of Space of Interest (SOI); 2) Two-Level Incremental Segmentation of SOI; 3) Determination of the final ROIs.

A. Generation of SOI

The aim of this step is to remove image regions that are not likely to contain pedestrians based on the knowledge of the geometrical structure of the scene, e.g. the road surface and the sky. By removing these regions in advance, we can significantly reduce the space to search for pedestrians. The remaining image regions refer to Space of Interest (SOI). Each ROI is an isolated candidate within the SOI. SOI is determined as follows: Ground plane is estimated using existing techniques proposed in [9, 10]. Distant regions are those whose disparity values are smaller than some predefined threshold. After removing the road surface and the distant regions, the remaining region in the image is the SOI as illustrated in Fig. 1 (c).

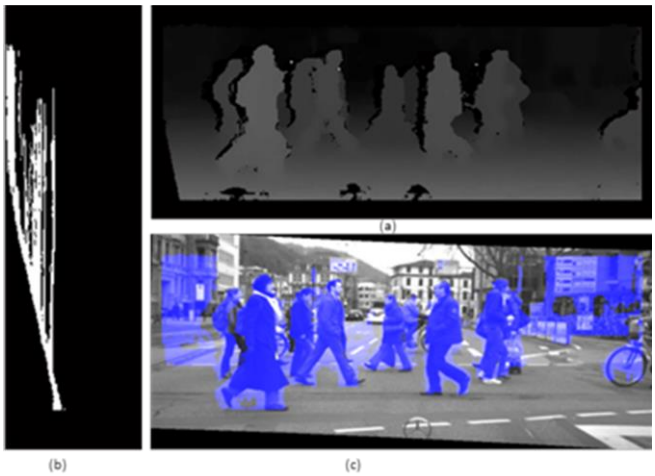

Figure 1. (a) a disparity map; (b) the corresponding v-disparity image; (c) space of interest highlighted in blue.

B. *Two-level Incremental Segmentation of SOI*

Segmentation of SOI into set of ROIs is performed on the u-disparity image. As explained in Section III, u-disparity image provides a bird-eye's view of the scene and the peak regions in the u-disparity image correspond to potential candidates for objects. These peak regions can be identified using connected component labeling technique. However, with the traditional 4-connectivity or 8-connectivity, multiple objects in close proximity are likely to be merged into one cluster as shown in Fig.2 (b). On the other hand, the more stringent 2-connectivity labeling is likely to partition an single object into multiple parts as shown in Fig. 2(c). The reason for this phenomenon is due to the fact that the fidelity

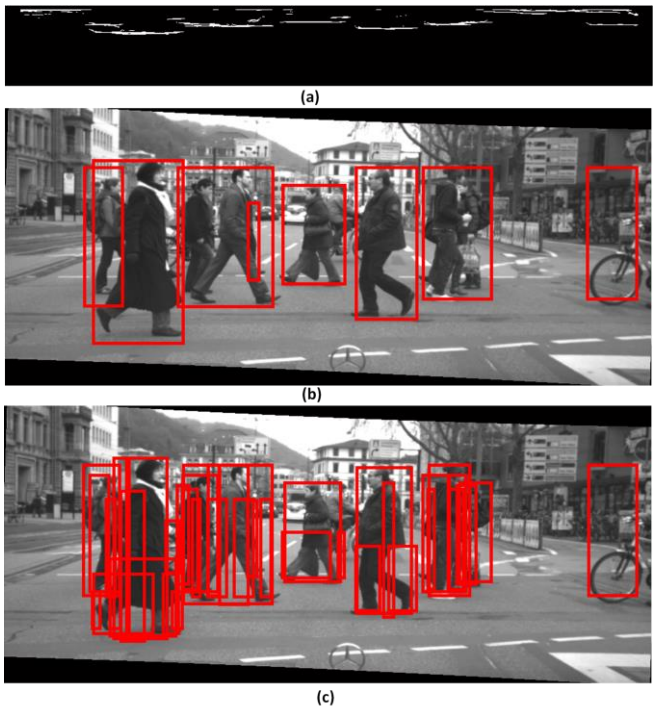

Figure 2. (a) u-disparity image; (b) ROIs generated using 8-connectivity connected component labeling for the near region; (c) ROIs generated using 2-connectivity connected component labeling for the near region.

of disparity map achieved by existing stereo matching algorithms is limited. In addition, according to (3), a minor change in d will resulting in a minor change in Z for near regions but a major change in Z for far regions.

The above analysis implies that varying connectivity must be employed for different peak regions when performing connected component labeling. The proposed method overcomes this problem with an adaptive strategy that allows the connectivity to change based on disparity values.

A two-level incremental segmentation strategy is proposed. The first level of segmentation employs an adaptive connected component labeling technique to segment a small set of clusters in SOI. This ensures that single objects are not wrongly partitioned into multiple clusters due to limited fidelity of disparity map. In the second level of segmentation, the clusters are then further refined to distinguish pedestrians in close proximity.

In the first level of segmentation, the disparity value with sub-pixel accuracy in the disparity map is directly rounded to the nearest integer. Based on this coarse resolution disparity map, the u-disparity image corresponding to the SOI is generated. The SOI is divided into two ranges: near range with disparity > *d_reference*, and far range with disparity <= *d_reference*. The hysteresis thresholding technique in [17] is applied to the u-disparity image to remove noise. The thresholds used in the far range are slightly lower than the ones in near range. Once the peak regions in the u-disparity image are identified, an adaptive connected component labeling technique is applied to perform clustering. Unlike the classical connected component labeling algorithm, which processes at the pixel level, the adaptive approach processes based on u-span. U-span refers to a continuous interval of peak regions in each row of the u-disparity image. Processing at the u-span level can lead to significant reduction in the clustering complexity. A u-span whose left-most position, right-most position and associated disparity value are *u_left*, *u_right*, and *u_d* respectively is denoted as *u-span[u_left, u_right, u_d]*. And *Rect[r_left, r_right, r_top, r_bottom]* denotes a rectangular region whose left-most, right-most, top-most and bottom-most position are *r_left*, *r_right*, *r_top* and *r_bottom* respectively. Then the examination neighborhood region of a *u-span[u_left, u_right, u_d]* is defined as *Rect[u_left-2, u_right +2, u_d, u_d + neighborhood_threshold]*. The value of *neighborhood_threshold* is determined adaptively based on (5), i.e.

$$k = \frac{u\_d}{d\_reference} \tag{6}$$

$$neighborhood\_threshold = \left\lceil \frac{k^2 * d\_reference}{d\_reference + 1 - k} \right\rceil \tag{7}$$

For a given u-span X, any u-span Y≠X which falls in the examination neighborhood region of X is assumed to be connected to X.

Figure 3(a) shows the results after the first level of segmentation. Since the relaxed connectivity compensates for the limited fidelity of disparity map, single objects cannot be easily divided into multiple clusters. However, a cluster may contain pedestrians who are in close proximity. This is due to

the fact that a minor change in disparity might lead to a large change in Z as explained in Section III. Therefore, we need to examine the clusters to check whether they consist of group of objects. This is achieved by using a fine-grained u-disparity image and stricter definition of connectivity.

In the second level of segmentation, the disparity map with sub-pixel accuracy, which corresponds to the SOI, is first multiplied by a factor and then rounded to the nearest integer. A corresponding fine grained u-disparity image is generated. As illustrated in Fig. 3(b), this operation achieves a 'zoom in' effect which amplifies the gaps between objects. A stricter connectivity is defined for the second level of segmentation. For a given *u-span[u_left, u_right, u_d]*, its examination neighborhood region is defined as *Rect[u_left, u_right, u_d, u_d+1]*. Each cluster that is obtained from the first level of segmentation will undergo a fine grained segmentation. This is achieved by applying the connected component labeling onto the corresponding region of the clusters in the fine grained u-disparity image with this new definition of connectivity. By doing so, clusters corresponding to single objects with good disparity representation will be retained, while clusters containing multiple objects will be further partitioned into new clusters. At the same time, single objects with gaps caused by erroneous disparity representation may also be further divided.

It is worthy to note that the aim of ROIs generation is not to extract the exact number of ROIs, where each ROI correspond to exactly one physical object. Rather the aim is to generate minimal ROIs without omitting any pedestrian. The coarse and fine grained segmentation strategies proposed above complement each other to generate a small set of ROIs. In addition, group of objects are separated and for each object, there is one ROI that corresponds to its full extent.

C. Determination of the final ROIs

Each cluster obtained from the earlier step corresponds to one ROI in the scene. The bounding box of the ROI is determined as follows: The left and right boundaries of ROI are determined by the left-most and right-most positions of the cluster in the u-disparity image. A typical way to determine the top and bottom boundaries of the bounding box is to rely on the vertical line in the v-disparity image. We argue that this operation can lead to inaccuracies since the height of a vertical line in the v-disparity image corresponds to the highest object at that distance (and not necessary the object of interest). When there are two objects with different heights at the same distance from the vehicle, the height of the shorter object will be wrongly determined since its vertical line in v-disparity image is occluded by the vertical line of the higher object. Instead, we determine the top and bottom boundaries of the bounding box by scanning from the row corresponding to the road in the disparity map, and identifying the regions whose disparity value is in the range of disparity values covered by the corresponding cluster.

Finally, post-processing is applied based on the size constraint and texture information. In this work, we assume the height range of a pedestrian in an image scene is 0.5 m (child) – 2 m (adult). ROIs will be removed if they lack of sufficient canny edge points and their size does not comply with the perspective effect on distance and the predefined pedestrian height.

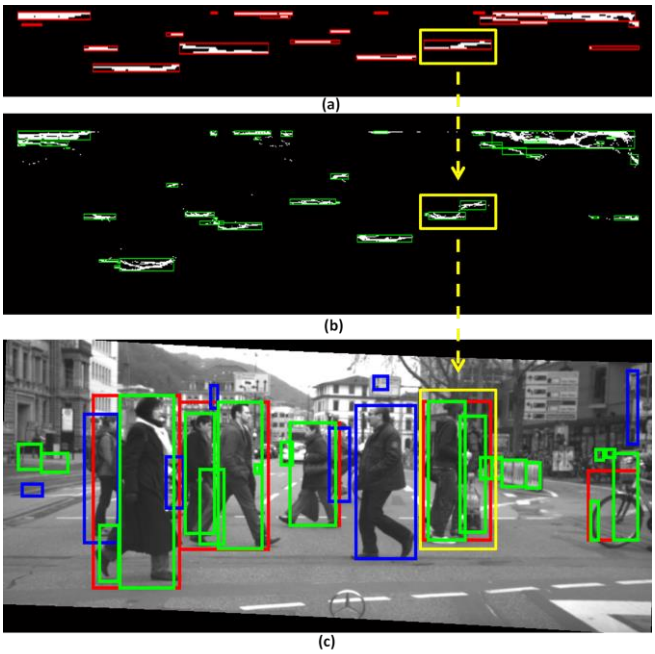

Figure 3. (a) clusters identified from the first level of segmentation in the coarse grained u-disparity image. (b) clusters identified from second level of segmentation in the fine grained u-disparity image. The fine grained u-disparity image is akin to the 'zoom in'effect of the coarse grained u-disparity image. (c) final ROIs generated using proposed method. ROIs in red are identified from the first level while ROIs in green are identified from the second level. ROIs in blue corresponds to the clusters identified from both levels. The yellow inset highlights an example where pedestrians are in close proximity and they can be distinguished by the proposed method.

V. Evaluation

We have chosen a publicly available benchmark [18] for evaluation. The dataset is taken using a stereo camera installed on a moving vehicle in a busy urban scenario. In this dataset, there are numerous cases where pedestrians are positioned close to other objects like pedestrian, vehicle, building, etc. The corresponding disparity map is computed using the stereo matching algorithm proposed in [16].

Two state-of-art works have been selected as the baseline algorithms. The first one is proposed in the well-known PROTECTOR project [6-8], which is denoted as *Baseline_A* in this paper. As mentioned in Section II, this method is a variant of sliding window scheme which tends to generate a large number of ROIs as depicted in Fig. 4(a). The second baseline, denoted as *Baseline_B*, is the work proposed in [9-12], which is based on the u-v disparity image framework. However, the work does not take full advantage of the disparity property and tends to merge group of objects which are in close proximity, as shown in Figure 4(b). The proposed method, on the other hand, fully exploits the property of disparity image and utilizes a two level incremental segmentation strategy. This results in only a small set of ROIs. In addition, group of objects are separated and for each object, there is one ROI that corresponds to its full extent. More qualitative comparisons between the two baseline

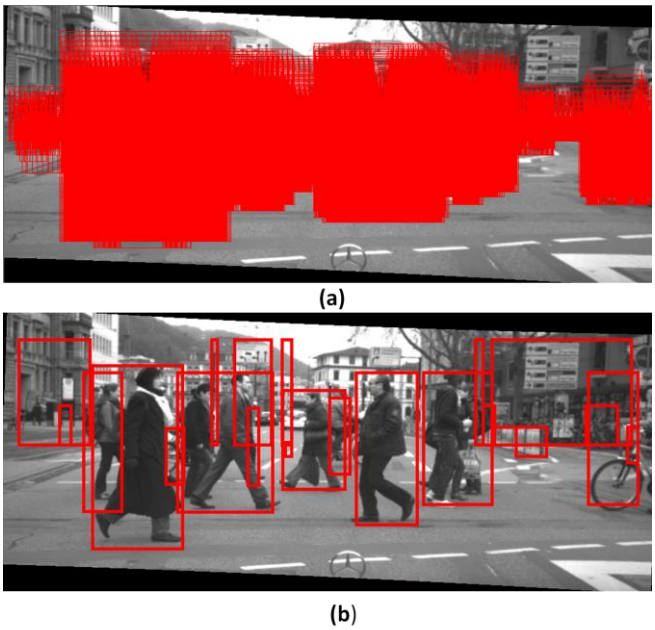

Figure 4. (a) Final ROIs generated using *Baseline_A*. The indistinct regions in red comprise of large number of overlapping ROIs. (b) Final ROIs generated using *Baseline_B*.

algorithms and the proposed method are shown in Figure 5. It is also worthy to note that the proposed algorithm is resilient to occlusion and sensitive to small sized pedestrian (e.g. a child).

We have also conducted a quantitative performance evaluation of the ROIs generation algorithms. The evaluation metrics proposed in [19] are adopted: 1) PC - the amount of ROIs generated; 2) TPR - the true positive rate. From Table 1, Baseline A obtains a good TPR at the cost of generating large number of ROIs, which will cause heavy computation burden in the subsequent classification module. Baseline B has a low TPR as it tends to merge group of pedestrians into a single one in busy urban scenes. In contrast, the proposed algorithm not only generates a small number of ROIs but also achieves a high detection rate.

The main reason that prevents the proposed algorithm from achieving 100% detection rate is due to pedestrians that are located at far distance (i.e. more than 30 meter) and are close to other objects. In such cases, due to the limited fidelity of disparity map, the far-away pedestrians and their neighboring objects are associated with the same disparity value, and hence they cannot be differentiated. However, these pedestrians are beyond the high-risk area and therefore do not pose problem in safety-related applications.

TABLE I. QUANTITATIVE PERFORMANCE EVALUATION

Algorithm	PC (per-frame)	TPR
Baseline_A	3041	100%
Baseline_B	14	52.96%
Proposed	20	73.52%

VI. CONCLUSION

We have shown with a challenging benchmark, that the proposed two-level incremental segmentation is capable of generating minimal ROIs that can distinguish pedestrians in close proximity. In addition, it does not omit pedestrians in the high-risk regions of vehicle safety applications. It achieves this by first applying an adaptive strategy to identify a minimal set of clusters from the coarse u-disparity image. This step ensures that single objects will not be wrongly segregated into multiple ROIs due to limited fidelity of disparity map. The second level of segmentation, which employs a strict connected component labeling approach, further refines the clusters to distinguish pedestrians who are close to one another. The proposed method is based on the u-v disparity image framework and does not require complex computations, thereby lending itself well towards cost effective realizations. Finally, the minimal number of ROIs generated will further reduce the overhead of pedestrian classification.

REFERENCES

[1] D. Geronimo., A. M. Lopez., M. Lopez., A. D. Sappa., and T. Graf., "Survey of Pedestrian Detection for Advanced Driver Assistance Systems," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, pp. 1239-1258, 2010.

[2] D. F. Llorca, M. A. Sotelo, A. M. Hell ń, A. Orellana, M. Gavil án, I. G. Daza, and A. G. Lorente, "Stereo regions-of-interest selection for pedestrian protection: A survey," *Transportation Research Part C: Emerging Technologies*, vol. 25, pp. 226-237, 2012.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition(CVPR), 2005*.

[4] C. Papageorgiou and T. Poggio, "A Trainable System for Object Detection," *International Journal of Computer Vision*, vol. 38, pp. 15-33, 2000.

[5] P. Sudowe and B. Leibe, "Efficient Use of Geometric Constraints for Sliding-Window Object Detection in Video," in *Computer Vision Systems*. 2011.

[6] D. Gavrila and S. Munder, "Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle," *International Journal of Computer Vision*, vol. 73, pp. 41-59, 2007.

[7] D. M. Gavrila, J. Giebel, and S. Munder, "Vision-based pedestrian detection: the PROTECTOR system," in *Intelligent Vehicles Symposium, 2004*.

[8] C. Keller, D. Llorca, and D. Gavrila, "Dense Stereo-Based ROI Generation for Pedestrian Detection Pattern Recognition." in *DAGM*, 2009.

[9] R. Labayrade, D. Aubert, and J. P. Tarel, "Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation," in *Intelligent Vehicle Symposium, 2002*.

[10] S. J. Krotosky and M. M. Trivedi, "On Color-, Infrared-, and Multimodal-Stereo Approaches to Pedestrian Detection," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 8, pp. 619-629, 2007.

[11] Z. Hu and K. Uchimura, "U-V-disparity: an efficient algorithm for stereovision based scene analysis," in *Intelligent Vehicles Symposium, 2005*.

[12] M. Bertozzi, E. Binelli, A. Broggi, and M. D. Rose, "Stereo Vision-based approaches for Pedestrian Detection," in *Computer Vision and Pattern Recognition - Workshops (CVPRW), 2005*.

[13] S. Nedevschi, S. Bota, and C. Tomiuc, "Stereo-Based Pedestrian Detection for Collision-Avoidance Applications," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 10, pp. 380-391, 2009.

[14] D. Fernandez, I. Parra, M. A. Sotelo, P. Revenga, S. Alvarez, and M. Gavilan, "3D Candidate Selection Method for Pedestrian Detection on Non-Planar Roads," in *Intelligent Vehicles Symposium, 2007*.

[15] I. P. Alonso, D. F. Llorca, M. A. Sotelo, L. M. Bergasa, T. Pedro Revenga de, J. Nuevo, M. Ocana, and M. A. G. Garrido,

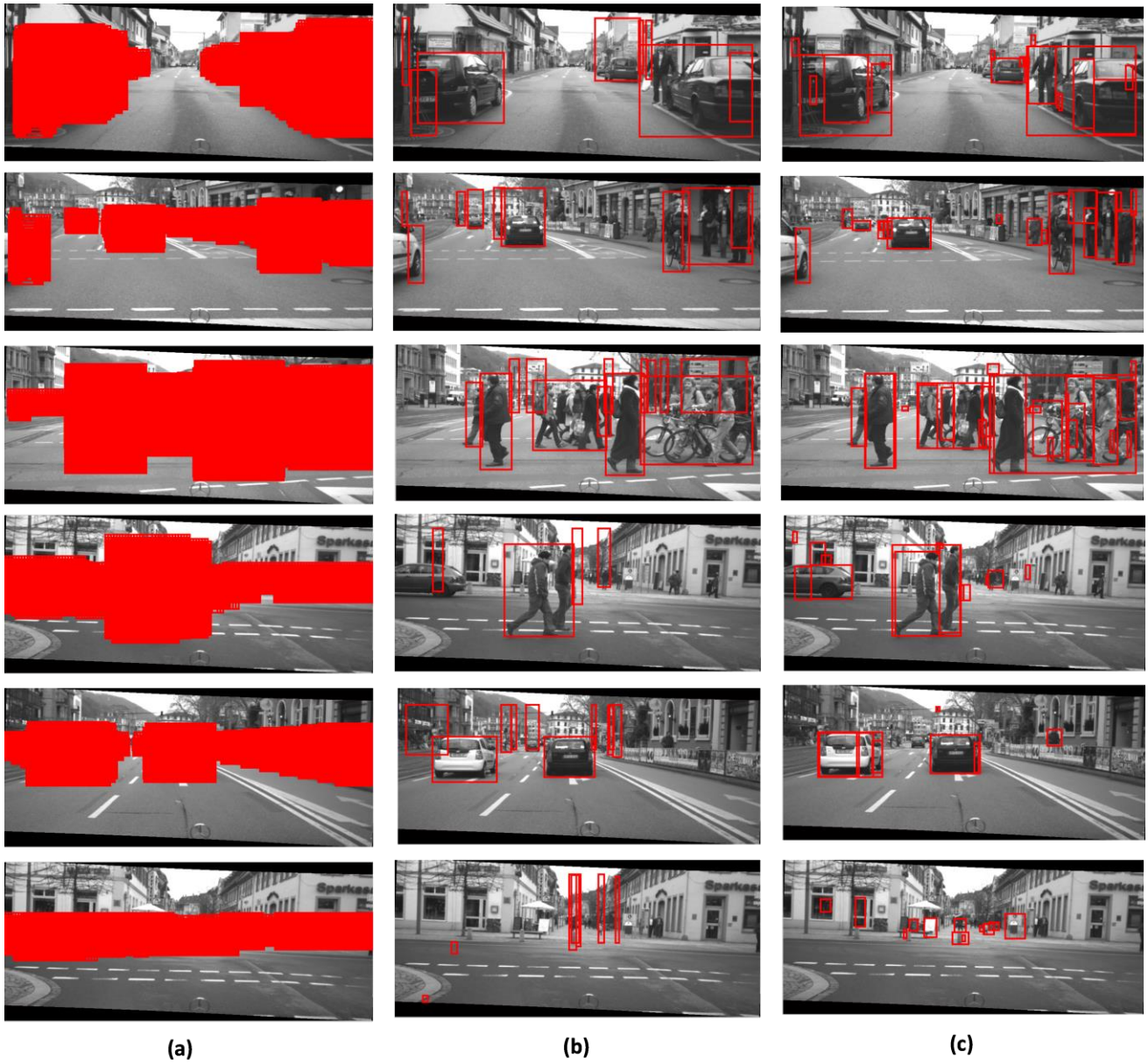

Figure 5. Qualitative comparison between the two baseline algorithms and the proposed method. (a) ROIs generated by *Baseline_A*; (b) ROIs generated by *Baseline_B*; (c) ROIs generated by proposed method

"Combination of Feature Extraction Methods for SVM Pedestrian Detection," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 8, pp. 292-307, 2007.

[16] H. Hirschmuller, "Stereo Processing by Semiglobal Matching and Mutual Information," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, pp. 328-341, 2008.

[17] J. Canny, "A Computational Approach to Edge Detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-8, pp. 679-698, 1986.

[18] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth, "Efficient Multi-cue Scene Segmentation," in *Pattern Recognition*, 2013.

[19] D. Cheda, D. Ponsa, and A. M. Lopez, "Pedestrian candidates generation using monocular cues," in *Intelligent Vehicles Symposium (IV), 2012* .